Accurate CMOS Compact Model and the Corresponding Circuit Simulation in the Presence of Statistical Variability and Ageing.

Jie Ding

Submitted to the University of Glasgow, School of Engineering

in fulfilment of the requirements for

the degree of Doctor of Philosophy.

September 2015

Abstract

As CMOS scales down to sub-50 nm, it faces critical dimensions of charge and matter granularities, leading to the drastic increase of device parameter dispersion, named statistical variability, which is one of the main contemporary challenges for further downscaling and makes each device atomistically different leading to broad dispersion of their electrical characteristics. In addition, device reliability concerns gain inertia; among them Bias Temperature Instability (BTI) shortens device lifetime by trapping charges in defect states of the insulator or at the interface. The interplay between statistical variability and BTI results in more variations on device performance and thus greatly affect circuit performance. In turn design methodologies must evolve towards variability and reliability aware design. To do so statistical compact models including both the effects of statistical variability and BTI-induced ageing are required for the large-scale statistical circuit simulation of variability and reliability.

In this study, the application of accurate compact models, that describe performance variation in the presence of both statistical variability and reliability at arbitrary BTI-induced ageing levels, to SRAM circuit simulation is described. Both SRAM cell stability and write performance are evaluated and it is seen that, due to the accurate description of device performance distributions provided by the compact models and the sensitivity of these SRAM performance metrics on device performance, the approach presented here is better suited to high-sigma statistical circuit analysis than conventional approaches based upon assumed Gaussian distributions. The approach is demonstrated using a 25 nm gate length bulk MOSFET whose performance variation is obtained from statistical TCAD simulation using the GSS simulator GARAND. The simulated performance data is then used directly as the target for BSIM4 compact model extraction that ensures device figures of merit are well resolved for each device in a statistical ensemble. The distribution of compact model parameters is then generalised into an algebraic form using Generalized Lambda Distribution (GLD) methods, so that a sufficiently large number of compact models can later be generated and interpolated at arbitrary ageing levels. Finally compact models generated in this way are used to

evaluate SRAM write performance and stability under the influence of statistical variability and BTI-induced ageing.

Acknowledgements

Firstly, I would like to give my sincere gratitude to my first supervisor: Prof. Asen Asenov, who has encyclopaedic knowledge of MOSFETs, led me to the device modelling world, and guided me with instructive advice throughout this study. I would like to thank my second supervisor: Dr. Dave Reid, who devoted numerous hours and patience. I am very grateful for his valuable comments and help. I would also like to thank Dr. Plamen Asenov for the technical discussions, and all the people in Device Modelling Group and Gold Standard Simulation for their help and encouragements.

A special thank to my parents for the endless love and forever support, I love you!

Finally, this work wouldn't have been possible without funding from James Watt scholarship and European project MORDRED.

Publications

Journal paper

IEEE Transactions on Electron Devices: **Jie Ding**, Dave Reid, Plamen Asenov, Campbell Millar, Asen Asenov. Influence of transistors with BTI induced ageing on SRAM write performance. 2015.

Microelectronics Reliability (Invited): L. Gerrer, **J. Ding**, S. M. Amoroso, F. Adamu-Lema, R. Hussin, D. Reid, C. Millar, A. Asenov. Modelling RTN and BTI in nanoscale MOSFETs from device to circuit: A review. 2013.

Peer Reviewed Conference Papers

ESSDERC 2013 (Talk): Jie Ding, Dave Reid, Campbell Millar, Asen Asenov Investigation of SRAM using BTI-Aware Statistical Compact Models. ESSDERC 2013.

SISPAD 2013 (Talk): Jie Ding, Dave Reid, Campbell Millar, Asen Asenov. An Accurate Compact Modelling Approach for Statistical Ageing and Reliability. SISPAD 2013.

VARI 2012 (Talk): Jie Ding, Plamen Asenov, Dave Reid, Campbell Millar, Asen Asenov. Statistical Compact Model Extraction in the Presence of BTI Degradation. VARI 2012.

IRPS 2013 (Talk): L. Gerrer, S. M. Amorose, P. Asenov, **J. Ding**, B. Cheng, F. Adamu-Lema, S. Markov, A. Asenov. Interplay Between Statistical Reliability and Variability: A Comprehensive Transistor-to-Circuit Simulation Technology.

SISPAD 2015 (Talk): R. Hussin, L. Gerrer, **J. Ding**, S. M. Amoroso, L. Wang, M. Simicic, P. Weckx, J. Franco, A. Vanderheyden, D. Vanhaeren, N. Horiguchi, B. Kaczer and A. Asenov. Reliability aware Simulation Flow: from TCAD Calibration to Circuit Level Analysis.

ESSDERC 2015 (Talk): Razaidi Hussin, Louis Gerrer, **Jie Ding**, Liping Wang, Salvatore M. Amoroso, Binjie Cheng, Dave Reid, Pieter Weckx, Marco Simicic, Jacopo Franco, Annelies Vanderheyden, Danielle Vanhaeren, Naoto Horiguchi, Ben Kaczer and Asen

Asenov. Statistical simulations of 6T-SRAM cell ageing using a reliability aware simulation flow.

ISCAS 2015: A. Asenov, **J. Ding**, D. Reid, P. Asenov, S. Amoroso, F. Adamu-Lema, L. Gerrer. Unified approach for simulation of statistical reliability in nanoscale CMOS transistors from devices to circuits

Workshops

PAnDA Workshop 2014 (Talk): Jie Ding, Dave Reid, Plamen Asenov, Campbell Millar, Asen Asenov. Evaluating the impact of ageing on SRAM stability using accurate statistical Compact Models.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

As Integrated Circuit (IC) continues to advance in terms of density, driven by the successful scaling of the transistors that they are built upon, they provide more and more functionality and the complexity greatly increases. For example, the current state of art 30 x 16.5 x 1.05 millimetre sized chip (Intel's new core M chip) contains billions of transistors. The unimaginable complexity forces circuit designers to use Computer Aided Design (CAD) tools to perform and optimise circuit design, and test circuit functionality while adhering to the restrictive design rules required by the manufacturing process. In addition, as the cost associated with fabrication of modern devices continues to increase due to the number and complexity of process steps required, design verification using CAD tools minimises the cost in early testing and speeds up the design process. Simulation Program with Integrated Circuit Emphasis (SPICE) is one of the most extensively used CAD tools. It deals with analogue, digital, as well as mixed digital and analogue circuits, analysing and solving circuit design problems. Since design defects can be spotted before production, SPICE greatly reduces the time, labor and cost instead of repeatedly producing test chips. By doing AC, DC and transient etc. analysis, SPICE effectively determines if the circuit realizes the expected functionalities and meets the performance targets. Meanwhile, circuit performance at ideal and variant non-ideal situations can be simulated, which is of great importance for design improvement.

SPICE circuit simulation is realized by connecting element models according to the circuit design. Transistor compact models (hereafter referred to as compact models) are vitally important, as they are the most basic functional element in any IC design. The accuracy of compact models is critical in determining the accuracy of subsequent circuit simulations. The compact models act as the bridge between the underlying technology and the integrated circuit design, providing quantitative insight into the relationship between device design choices and circuit performance. Only with advanced and accurate compact models can SPICE simulation become a tool to drive towards competitive designs. Various compact models emerge with various SPICE software. The Compact Model Council (CMC) was founded with the aim of standardizing compact models and their parameters in order to make them compatible in different simulators. BSIM4 is the industrial standard compact model for bulk Complementary MOS (CMOS), standardized in 2000 by the Compact Model Council [1, 2]. BSIM4 is exclusively used in this work.

As CMOS technology scales down into the sub-50-nm regime, device modelling becomes tremendously challenging. Complex 3D structures, quantum mechanical confinement effects, the introduction of novel materials and stress engineering must all be addressed within accurate device simulation. However, accurate circuit simulation must go beyond the accurate simulation of an ideal device. There are two important effects that must be accounted for, that of intrinsic statistical variability and BTI-related reliability.

Within the ultra-small ‘atomistic’ device regime, statistical variability arising from discreteness of charge and granularity of material becomes a significant concern [3]. Aggressive scaling makes the device ultra-sensitive to the random fluctuations induced by statistical variability and results in initial characteristics different in each device [4]. The important sources of variability are the number and position of individual impurity atoms within the device, the lithographic irregularities that produce variation in physical gate length, and the granularity of metal gate that affects the gate workfunction. Statistical variability is an entirely stochastic phenomenon and is thus impossible to suppress at a technology (manufacturing) level. This means that within an IC design with billions of transistors, an accurate assessment of the circuit design must reflect the fact

that each device will not have a single, well characterised performance, but will instead be from a distribution of performances.

The introduction of high-k dielectrics together with metal gates [5] in order to allow further scaling of Equivalent Oxide Thickness (EOT) while limiting gate leakage, comes at the expense of transistor reliability, mainly Bias Temperature Instability (BTI), which induces “ageing” problem [6]. Because the High-K material cannot be well controlled, the trapping of electrons and/or holes at the interface and in the gate stack during circuit operation is unavoidable and is cumulative over time. This accumulated damage with time is the reason behind the term “ageing”. The age related degradation has recoverable and permanent components that seriously affect device threshold voltage and result in circuit failure.

Moreover, statistical variability and BTI can interact with each other, making the problem more complicated and variations more severe. This leads to a great necessity to integrate statistical variability and BTI-induced ageing into compact model.

Worst-case circuit analysis, in which compact models under the worst conditions are applied, is the traditional method for predicting circuit behavior. If the corresponding worst-case circuit can still realize the required functionality under such an extreme condition, it should work well at the normal condition. However, without an accurate understanding of the statistical nature of the device variation, a pessimistic worst case is typically adopted to guard against the unknown. This necessarily results in pessimistic circuit performance prediction. Statistical variability and BTI makes each transistor microscopically different and thus makes the circuit perform differently with different transistors. Therefore, such deterministic design flow needs to be shifted toward statistical circuit design flow. To do this, accurate and reliable compact models that can represent the statistical properties of device performance under the influence of statistical variability and BTI-induced ageing are required in order to ensure the accuracy of circuit simulations. Therefore, propagating statistical variability and ageing effects into compact models is necessary and important for the purpose of design verification, as well as power/performance/yield predictions (PPY).

Static Random Access Memory (SRAM) plays an important role in microprocessor design and is always used as the benchmark of a technology node. It is array arranged into into the high density memory, containing a number of minimum dimensioned transistors. Therefore, strict design constraints and power density for SRAM are applied on chip area. SRAM functionality requires well balanced transistors that makes SRAM cells very vulnerable to the mismatch induced by statistical variability and BTI. And the failure of a single SRAM cell can be catastrophic. In this research, SRAM is used as the workhorse to investigate the influence of statistical variability and BTI-induced ageing at circuit level.

1.2 Aims and Objectives

Until now little research has been done to integrate statistical variability and BTI-induced ageing effects together into compact models and investigate their effects at circuit level. Therefore, the ultimate aim of this PhD thesis is to integrate the impact of statistical variability and reliability into compact models and investigate the influence on SRAM performance in advanced emerging CMOS technology. In order to achieve this main aim, the work should achieve the following objectives:

1. Obtain and analyse device performance data in the presence of statistical variability and at different levels of BTI-induced ageing. This data, and the trends discovered therein will then be used to inform our compact model extraction.
2. Develop the details of a physically based statistical compact model extraction strategy, and assess the accuracy of that extraction on a realistic ensemble of bulk 25 nm devices. Accuracy should be assessed against device figures of merit and their correlations at different ageing levels.
3. Develop and test a method of generating additional compact models at arbitrary ageing levels using the information of the physically extracted compact models, to facilitate later circuit simulation. The generated compact models should be assessed against the device figures of merit and their correlations under the influence of statistical variability and BTI-induced ageing compared with the original device performance data.
4. Investigate the impact of statistical variability and BTI-induced ageing at circuit level, using the results of objectives 1-3. In this case, we will use a 6-T SRAM as

a case study and will investigate the stability and write performance of the SRAM cell.)

1.3 Outline

The work in this thesis is organised in seven chapters. Chapter 1 is the introduction. Chapter 2 is focused on the research background. Chapter 3 introduces the methodologies used in this research. Chapters 4 shows the physical simulation and extracted compact model results, which achieve objective 1 and 2. Chapter 5 introduces GLD and interpolation methods. Using these methods, compact models can be generated at arbitrary ageing level. Chapter 6 mainly uses the generated compact model on SRAM circuit to investigate the influence of statistical variability and BTI-induced ageing at circuit level. Chapter 7 is the final chapter, in which conclusions are drawn and future work is discussed.

Chapter 2 starts with the development of CMOS technology and introduces the process of transistor scaling down. Problems associated with aggressive scaling, statistical variability and BTI, are introduced in section 2.2, 2.3 and 2.4, including variability classification, statistical variability and BTI mechanisms, as well as the research background in these areas. Compact model development and up to date research in the field have been demonstrated in section 2.5. Finally, section 2.6 describes 6T SRAM circuit.

The main aim of Chapter 3 is to give a complete description of the methodologies used in this research. In section 3.2, the research flow and simulation tool chain are introduced. Physical simulation is used to obtain the device performance data in presence of statistical variability and BTI-induced ageing, and the methodologies used in physical simulation is introduced in section 3.3. Section 3.4 describes the two-stage compact model extraction methodology and the extraction strategy that is used in this research to extract compact models. In section 3.5, the Monde Carlo circuit simulation engine RandomSpice, used in this research for statistical circuit simulation is introduced.

Chapter 4 mainly shows the achieved results of objectives 1 and 2 by using the approaches in Chapter 3. It starts with the demonstration of physical simulation scenarios

with statistical variability and at different BTI-induced ageing levels. Physical simulation results are analysed in section 4.3. With the two-stage extraction method and the selected parameters introduced in section 3.4, compact models are extracted from physical simulations and the extracted models are compared with physical simulations, verifying the extraction accuracy.

Chapter 5 discusses the compact model generation methodology and results. Compact model generation aims at generating a sufficiently large ensemble of statistically accurate compact models at arbitrary trap densities. The reason for doing so is illustrated in section 5.1 and 5.2, showing the importance and necessity of compact model generation to avoid subsampling problems. Section 5.3 shows the inaccuracy of the extensively used Gaussian V_T generation method and introduces GLD, which is a new generation method. Also described is an interpolation method that enables compact model to be generated at arbitrary trap density is introduced in section 5.4. The ageing model that can translate between trap density and ageing time is introduced in section 5.5. Section 5.6 integrates GLD, interpolation and ageing model into RandomSpice. The accuracy of the method developed in this research is verified by comparing with physical simulation results in section 5.7.

In Chapter 6, the generated compact models are applied to 6T SRAM, for the investigation of the influence of statistical variability and BTI-induced ageing at circuit level. The SRAM metrics, Static Noise Margin (SNM) and dynamic Write Margin (WM) are investigated at different ageing scenarios.

Chapter 7 is the conclusion and future work discussion part. In this chapter, the research in this work is summarised and the conclusion is drawn. Meanwhile, the future possible work is discussed.

Chapter 2

Background

2.1 MOSFET Scaling

Since the first integrated circuit was realized by Jack Kilby in 1958 [7], the semiconductor industry has experienced exponential growth and tremendous success over 5 decades. Today, billions of transistors can be produced on a fingernail-sized [8]. These levels of electronic component integration are of great significance, allowing multiple system functions to be realized within a small physical area, and thus enabling the current high performance and low cost electronic devices and consumer products. Transistors are to integrated circuits what bricks are to houses. Therefore, transistor scaling and performance improvements make key contributions to low cost and high performance semiconductor electronics [9].

The famous 1965 Moore's law, which posits that the number of transistors on an integrated circuit doubles every 18 months [10], has driven the semiconductor industry through rapid growth for more than four decades. It implies that at each new technology generation, transistor feature sizes are reduced to 0.7 of their previous dimensions and a roughly two times increase in transistor density is achieved. Higher device density engenders an increase in system functionality for each chip design [11, 12]. In addition, it lowers manufacturing costs significantly - for the smaller the transistors, the more chips can be produced on a single wafer, which in Moore's own words results in "unit cost falling as the number of components per circuit rises". Therefore, the cost of one transistor has dropped more than ten thousand times since 1960 (Fig.2.1) and the trend still continues. During the years of 'happy scaling', transistor sizes were reduced without

changing transistor architecture, as lithographic processes improved. Simultaneously, transistor performance increased without big challenges in underlying device physics. Yield could be well controlled. All of these factors brought the semiconductor industry enormous advantages. However, although the IC industry enjoys the benefits that scaling has brought for 4 decades, the 'happy scaling' years are over. At the beginning of the 21st century, CMOS enters the post-Moore's law period and device dimensions are at a sub-100 nm scale. Problems arise due to aggressive scaling (Fig.2.2) and make miniaturization more and more difficult [11, 13, 14].

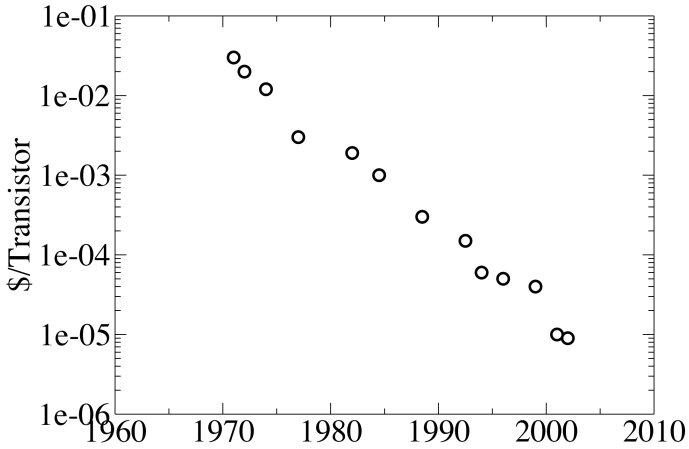


Fig.2.1 Average transistor price from 1968 to 2002. After [15].

For example, as the channel length is shrunk aggressively, poorer control of threshold voltage roll-off and drain-induced barrier lowering (DIBL) brings more variations to the threshold voltage and greatly increases average leakage current [16]. While scaling of the power supply voltage reduces the dynamic power consumption, the reduction of threshold voltage leads to larger subthreshold leakage. The exponential increase in IC density has been accompanied by an exponential increase in heat generation [17]. Moreover, and from the material perspective, SiO_2 gate dielectrics suffer from charge tunnelling and material breakdown issues, as they become thinner and thinner [18]. High-permittivity (high-k) materials were introduced as the gate dielectrics at the 45 nm technology generation [5, 19], allowing for a larger physical dielectric thickness, and leading to a reduction of leakage current and faster switching speed. However, for transistors below 50 nm, device scaling does not bring as much improvement as before due to the smaller dimension and area difference between the conjoint generations. On top of this, a significant investment in fabrication plant is required for the additional

complex process steps involved in making architectural changes such as the introduction of high-k dielectrics.

The semiconductor industry faces many challenges in order to continue the deep transistor scaling. In addition, it is desirable that each scaled transistors' characteristics are identical to the target characteristics (i.e. the same threshold voltage, drive current etc.), to ensure that CMOS integrated circuits employing hundreds of millions of transistors can realize complex functions reliably. For example, transistors with different threshold voltages but driven from a single ramping gate voltage, will have different turn-on times, and thus logic gates constructed from them will have different delay times, a source of race conditions in sequential logic. However, when transistors become smaller, small physical variations between transistors can result in large differences in their performance. For example, varying the width of a device ±5 nm on a 20 or 200 nm wide device, produces a 25% or 2.5% variation in drive current. A 5 nm variation in a transistor with nominal channel length of 20 nm results in a much higher deviation in device threshold voltage compared to a device with a nominal 200 nm channel length transistor. Variations between transistors make each transistor microscopically different, leading to the loss of yield. This becomes a serious obstacle for further transistor scaling.

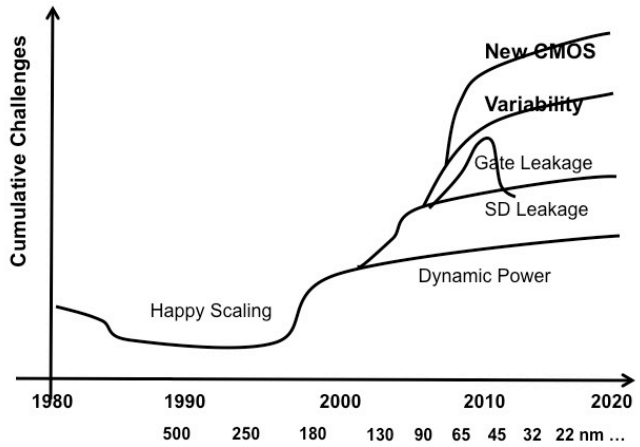


Fig.2.2 Cumulative interdependent challenges as a function of time (and technology generation). After [15].

2.2 Variability Classification and Statistical Variability.

Since variability has a crucial impact on transistor miniaturization, it has been the subject of extensive investigations [20-22]. Variability is characterized by its origin and behaviour [23]. We can classify it into two categories: 'process' induced variability and purely 'statistical' variability. Process variability includes wafer-to-wafer level, wafer level, die level and layout-dependent variations (Fig.2.3). Wafer-to-wafer level refers to variations between different wafers, normally caused by variations in processing machine and environment (humidity, temperature, pressure…) conditions. Wafer level is the on-wafer non-uniformities caused by gradual changes in temperature, gas flow, implantation doses etc. across the wafer as a whole. Variations on die level are often due to lithography steps of die-by-die pattern exposure. Layout-dependent variations are caused by the layout of patterns, such as pattern-to-pattern distance and pattern densities. For such 'process' variability simulations of the parameters drift of circuit components is typically amenable to modelling and characterization based on process and design parameters. Its effects are typically predictable and deterministic for any given circuit design. Therefore, normally such variability can be handled by improved process control or more regularity of the designed circuit layout.

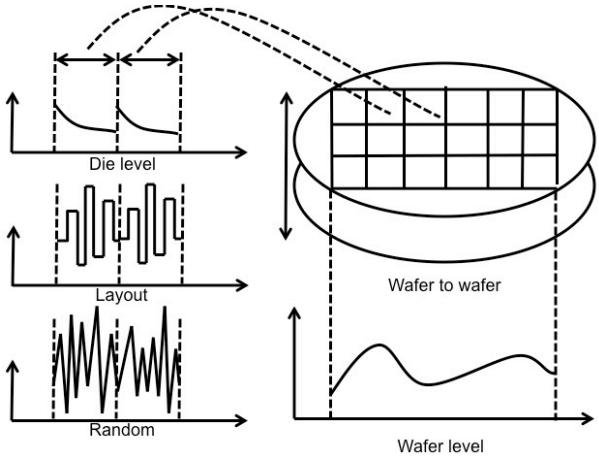


Fig.2.3 Classification of variability. After [23].

We use 'statistical' variability as a term for variability which is impossible to suppress purely at the technology (manufacturing) level. Such variability arises from the discrete

nature of charge and granularity of materials. It is truly stochastic and consequently, much harder to deal with – for it is impossible to predict the precise mismatch of any two components in a given circuit, but only the statistical likelihood of any particular mismatch. Though statistical variability has been investigated since 1975 [21], it did not bring the industry great concern for a long time. However, when devices began approaching deep submicron dimensions, statistical variability started to become a very critical issue [24]. It becomes more and more severe as device dimensions shrink further. For a long channel device, statistical variability induced variation has little impact on device performance as the statistical variations self-average over the relatively large dimensions and are overwhelmed by process variability. But when the geometrical dimensions drastically shrink, device dimensions are at the atomic level and statistical variability cannot be ignored any longer. For example, doping concentration is used as a single parameter to quantify the number of dopants in the channel. For larger devices, this is reasonable because even a difference of hundreds of dopants between devices will not result in a large difference in macroscopic device performance, considering the much larger number of total dopants in the channel. But when the channel length is under 50 nm (here we assume that the device is a minimally widthed square device), the total number of dopants in the channel is of the order of 10 to 100 [25]. Few more or few less individual dopant charges the dopant density of 1-10%, which can result in a measurable difference in the transistor's properties and cannot be ignored. Moreover, the specific positions of the dopants in the channel also have an important role in determining device performance. The traditional concept of concentration, which assumes a smooth, continuous distribution may no longer be applicable. A 1 nm thick dielectric layer consists of only 5 atomic layers, which means a least 20% variation can be induced because of atomic scale non-uniformity.

From the above we can see that statistical variability arises due to intrinsic variations at the atomic level. Statistical variability always exists regardless of device dimensions, but was not problematic until device dimensions approached the atomic scale. Since there is less self-averaging of variations when the transistors approach the atomic scale, statistical variability now represents a major stumbling block for continued scaling.

Compared with process variability, statistical variability is less controllable. It was reported that the influence of statistical variability exceed that of process variability at the 130 nm technology node [26]. Wafer-to-wafer or die-to-die variation induced by process variability can be improved or mitigated by design or process changes. However, statistical variability is induced by atomic level disorder, and cannot be avoided. It makes each transistor both microscopically and macroscopically different, which is contrary to the designer's expectation that identical devices will have the same electrical characteristics. This will introduce mismatch in circuits and will have a significant impact on the functionality, yield and reliability of the corresponding circuits and systems. The study of statistical variability is therefore critical to first understand, and then mitigate, the impact on circuit and system performance and yield.

In the following subsections, three main sources of statistical variability in bulk MOSFETs will be discussed. They are random discrete dopants (RDD), line edge roughness (LER) and metal gate granularity (MGG) [27-30].

2.2.1 Random Discrete Dopants

RDD are the dominant source of statistical variability in bulk MOSFET. When transistor dimensions are much larger than the atomic scale, device behavior can be simulated with the assumption that devices are continuously doped. Thus the carrier distributions and potential profiles are all smooth. However, as the dimensions of transistors approaches the atomistic scale, the number of dopants in the active region is of the order of 10 to 100 and the assumption that transistors are continuously doped is not reasonable any more. The random number and position of the individual dopants result in inhomogeneity in the potential profile. Fig.2.4 shows an example comparison between the potential profiles of continuously doped and atomistic devices. The inhomogeneous potential profile leads to local variation in the height of the source-to-drain barrier, as can be seen in Fig.2.4 (b), meaning that some regions of the device will turn on at different gate voltages compares to other regions, thus affecting the overall threshold voltage.

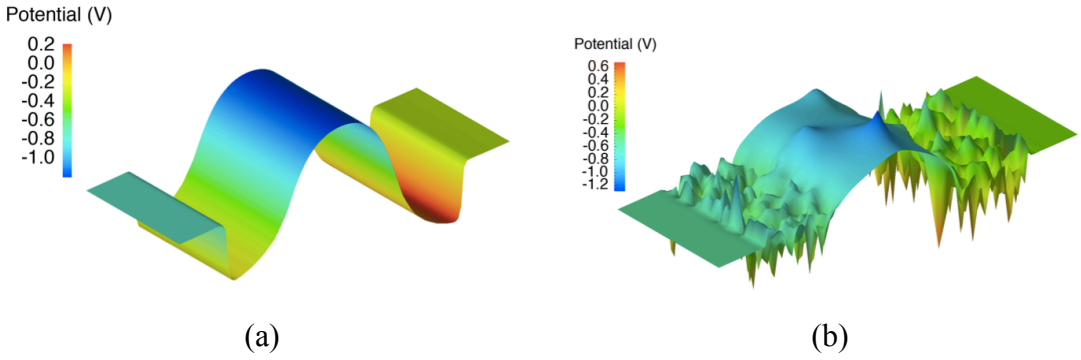


Fig.2.4 (a) Continuously doped device potential profile. (b) An atomistic device potential profile [31]. Used with permission.

As dopants are implanted into the device at high energy, they scatter many times before rest. During the thermal annealing process, the implanted dopants diffuse and replace the Si atoms, becoming electrically active. In the whole process, dopants are randomly scattered and diffused. Thus the dopant distribution is stochastic and varies from device to device, resulting in the variations in the nominal threshold voltage and hence mismatch between nominally identical devices. Fig.2.5 clearly shows that as the channel area decreases, the fluctuations in the threshold voltage greatly increase.

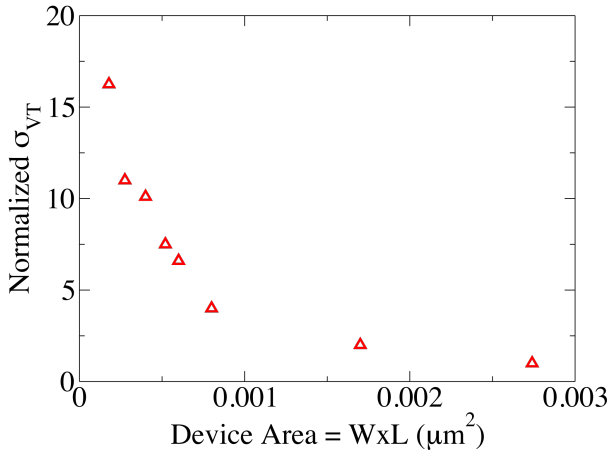


Fig.2.5 The deviation of threshold voltage as a function of channel area. After [32].

RDD was pointed out as early as in 1972 [14]. After this, experimental data showed the effect of RDD in many studies [33-35] from the 1980s. As MOSFETs have shrunk, RDD has been more extensively studied, from the analysis of the experimental results [33-37], to the modelling of its effect for the optimization of the device and circuit design [38-40]. The analytical models that describe threshold voltage fluctuations induced by the random

dopants in MOSFETs have developed from 2D [36, 41, 42] to 3D [39]. As has been shown in [43], RDD has contributed 60-65% of the total variability for 65 and 45 nm technologies. It is of great importance to consider RDD in conventional MOSFET design.

2.2.2 Line Edge Roughness

Though RDD is the dominant source of statistical variability in conventional planar MOSFETs, in the deep sub-micron regime another source of statistical variability begins to have a strong effect on device performance – namely, line edge roughness (LER) – which affects the precision with which the device can be patterned.

LER arises due to the granular nature of the resist polymer and the discrete nature of high absorption photon to photon. During the fabrication process, photoresist is spin-coated on the silicon fabricate wafers. Then the wafers are exposed to UV or deep UV light through the photomask, as a result of which the gate image is patterned on the photoresist. The exposed wafer is emerged in developer for a certain time according to the photoresist material and the corresponding areas are dissolved depending on the type of the resist. If the positive photoresist is used, the exposed area will be left after developing. Contrarily, the inverse area will be left if the negative photoresist is applied. However, during the exposure process, light diffraction or e-beam scattering leads to roughness in the edges of the exposed area. Meanwhile, the resist molecules entangle and polymer aggregates form. Roughness also arises when in dissolving the exposed photoresist, where larger aggregates take longer to dissolve than the smaller ones. Thus, the gate length varies microscopically across the width of the device. Subsequent processing duplicates the rough lines into the transistor structure by etching, and then the remaining photoresist is stripped.

LER is not a problem if the transistor's channel length is orders of magnitude larger than the RMS roughness. As reported in [44], LER has no effect on Intel's 130 nm devices because the channel length is much larger than the RMS roughness. Due to the negligible influence of LER on device performance during that period, modelling of its effect was nearly absent. However, with the transistor scaling, LER has become a significant

fraction of the gate length, introducing noticeable local variations in the gate length. It is reported that for the 45 nm technology that RMS of LER is almost 10% of its nominal length and it varies the saturation current as much as 10% [45]. As LER exerts more and more influence on device behavior, it becomes more and more important to be included into device simulator in order to properly assess its impact on device and circuits performances. The modelling and simulation of LER has changed from the approximation of using a 'square wave' [46, 47], as well as the simplified 2D simulations [48, 49] at the beginning, to the advanced statistical 3D simulation [50] to emulate the 3D stochastic effect of LER.

As gate length varies across the width of the device, device performance varies. For the longer parts of the gate, it is harder to turn on compared with the nominal gate length. The corresponding leakage current and drive current degrade. For the shorter parts of the gate, though it is easier to turn on and the drive current increases, the leakage current greatly increases. Due to the random nature of LER, the average device gate length is different from one device to the next. Therefore, the corresponding device threshold voltage, as well as leakage current and drive current vary from one transistor to another. These variations brought about by LER are very channel length dependent, being characterized by the V_{TH} roll-off characteristic of the device as a function of the average channel length.

2.2.3 Metal Gate Granularity

At the 45 nm technology, Intel changed the SiO_2 gate oxide to a high-k dielectric material for higher performance and lower leakage. Scaling the thickness of the dielectric layer is the main factor in the process of downscaling because it increases the capacitance between the gate and the channel and hence the mobile change in the channel. Thus with the same gate bias, improved gate dielectric scaling means that the inversion layer can be formed more quickly, and transistor switching speeds up. SiO_2 was abandoned because its thickness had reached the physical quantum mechanical tunnelling limit. For very thin gates (<1 nm) the leakage due to tunnelling through the gate oxide was becoming critical. Further scaling of SiO_2 thickness was therefore, impossible. In order to increase device performance and switching speed and minimise gate leakage, high-k materials were

introduced to allow the use of a thicker physical gate oxide while retaining the electrical equivalent thickness of the old, thin SiO_2 [5]. The corresponding polysilicon gate has been changed to metal gate for compatibility. This is called High-k Metal Gate (HKMG) technology [19].

However, the metal gate, which is deposited before dopant implantation and acts as a mask for the source and drain (gate-first technology), crystallizes after thermal annealing. As the gate becomes polycrystalline, grains with random sizes and different crystallographic orientations and workfunctions form. The granulized metal gate results in randomly distributed local threshold voltage variations in the gate region. Therefore, MGG arises as the third main source of statistical variability. Fig.2.6 shows the surface electrostatic potential of a 30x30 nm MOSFET, which has two grains in the gate with the boundary between grains diagonally across the gate [51]. The figure shows the local variation in the source-to-drain barrier dependent on the work function of the grain above that part of the channel.

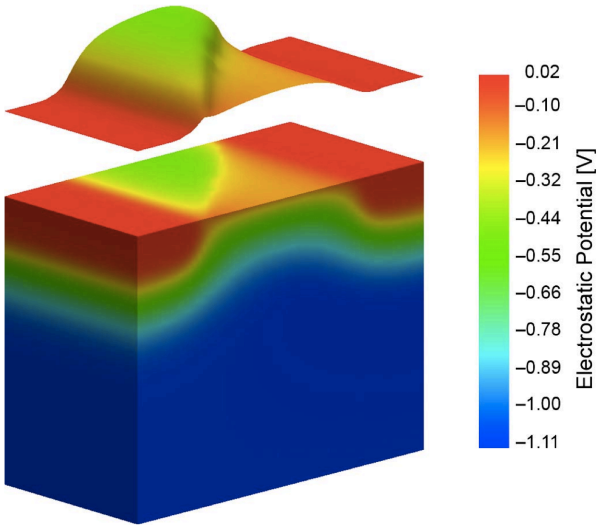


Fig.2.6 Electrostatic potential of metal gate MOSFET with two grains [51]. Used with permission.

MGG can be minimized by using a gate-last process, in which a sacrificial gate is used for dopant implantation and polished afterwards by Chemical/Mechanical Polishing (CMP) and removed by etching. The real gate is deposited after thermal annealing, remaining amorphous, and thus reducing the variations. However, in addition to the extra cost due to more steps in gate-last process, the surface after CMP can exhibit dishing and

irregularities, requiring further remedial processing steps. Therefore, there is a trade-off between using gate-first and gate-last process, with both being used in the semiconductor industry.

The modelling of the effects of MGG has been researched previously [51-55]. The proposed model in [53] considers the statistics of grain sizes and orientations. However, the relationship between grain size and gate size is not considered in this study. The equivalent model is also used in [54]. Afterwards, a full-scale 3D simulation is performed which considers both grain size and gate size and is based on a 30 nm gate length MOSFET [51].

With the above description of the three main sources of statistical variability, it is easier to understand the stochastic nature of statistical variability. A detailed qualitative and quantitative understanding of the sources of statistical variability will lead to a better analysis of stochastic device performance and corresponding circuit performance. Since some of these effects are non-Gaussian, and the important information for yield is in the tails of the distributions, large scale statistical analysis is essential at device and circuit level respectively for the investigation of the influence of statistical variability.

2.3 Bias Temperature Instability (BTI) induced Ageing

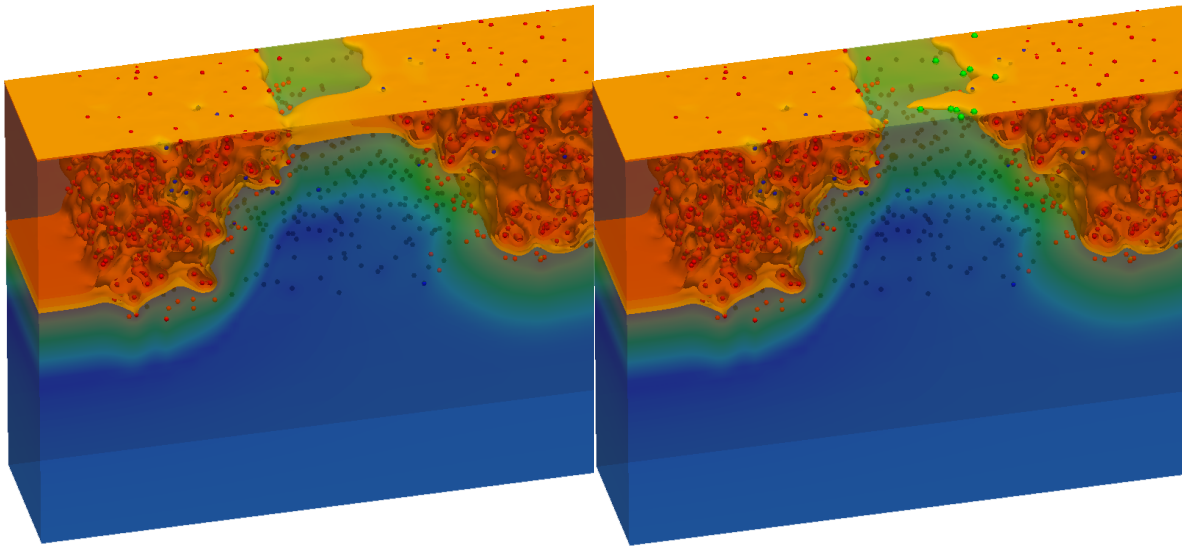
Statistical variability is a type of variation that is independent of time, but is determined by the discreteness of charge and granularity of matter. Once the device is produced, its characteristics will not change over time. BTI is another phenomenon that affects the device during their lifetime, and becomes more severe as time passes, resulting in device degradation. NBTI, which stands for Negative Bias Temperature Instability, refers to the negative bias applied to the gate. This was first observed as early as 1967 in PMOS. PBTI, which stands for Positive Bias Temperature instability, is not important in poly-gate NMOS. However, with the introduction of HKMG technology, NBTI in PMOS and PBTI in NMOS have become equally important [56].

BTI is caused by electrical and thermal stress. It results from electrons or holes trapped in defect states at the interface on the gate oxide during circuit operation, when transistors in the circuit are operated in a stress condition (e.g. high gate bias and elevated temperature) [57]. Some carriers become permanently trapped over time, resulting in a permanent additional fixed change. As stress time accumulates, the number of traps increases and the dielectric layer's insulating quality decreases, resulting in increased electron tunnelling through the dielectric. Therefore, as the device degrades the corresponding device threshold voltage shifts and drive current decreases [58, 59]. The effects of BTI become more severe as transistor dimensions shrink due to two reasons. Firstly, during production, high-k material has more defects than SiO_2, which amplifies BTI problems. Some research shows that PBTI appears in NMOS after HKMG and it is assumed due to the larger number of defects in the dielectric layer that trap more charges. However, the mechanisms behind PBTI are still not fully understood. Secondly, as devices have scaled, supply voltages have reduced to around the 1 V level, so the shift in V_{TH} due to BTI has an increasing impact on the circuit performance. The random number and position of the trapped charges also results in increased statistical variability.

Trapped charge density increases with device ageing. The effect of these trapped charges cannot be considered alone as their effect is inextricably linked to the other sources of statistical variability, and in particular to the other discrete dopants [60]. An additional stress-induced charge will have a large impact if it occurs along a favoured current percolation path where there are no discrete dopants, but will have little effect if it occurs within a cluster of dopants which already exclude current flow from that region of the channel. This is clearly demonstrated in Fig.2.7, where the fresh transistor has a current percolation path between source and drain. Three strategically trapped electrons result in a blocking of the current percolation path, producing a large increase of 145 mV in the threshold voltage. Therefore, BTI can not be considered independently of microscopic variations in transistor structure [59]. Investigation and simulation of BTI combined with statistical variability is necessary.

BTI is one of the most important device reliability concerns in advanced technologies [61, 62]. Threshold voltage shifts result in device operating characteristic drift, which changes over the lifetime of the devices, and can lead to circuit operation failure if it

exceeds circuit tolerance [63]. In small devices, when the intrinsic interaction of BTI with statistical variability is included, the problems can become severe. Different device turn on time due to device V_{TH} shifts brings in sub-circuits frequency degradation, and finally result in system failure. Therefore, the investigation of BTI associated with statistical variability is very important from both device to circuit point of view [64].



(a) (b)

Fig.2.7 The electron density in a single atomistically simulated device. In (a) a clear current path exists between the source and drain. (b) shows the effect of trapped charges (shown in green) which have completely closed off the current path resulting in a large change in device threshold voltage.

2.4 Compact Modelling

SPICE (Simulation Program with Integrated Circuit Emphasis) is the industry standard circuit simulation tool that is extensively used to assist circuit design. As the IC industry grows exponentially, circuit design can only be accomplished with a circuit simulator, since large simulations need to be carried out for function realization and design verification. In addition, due to the small feature size, complicated layout and manufacturing process, IC manufacture is costly. Thus, avoiding design defects and ensuring the designed circuit performs as expected before production is of great importance. SPICE plays a vital role in facilitating circuit design, verifying circuit performance and predicting yield. As well as being used for existing technologies,

SPICE is also used for the early-stage design before the new technologies are formally introduced. In order to guarantee SPICE simulation accuracy, electrical component models in the circuit should mimic reality as closely as possible. On the other hand, these models should be as simple as possible to efficiently simulate the circuit, especially for circuits with many components. These simplified models, which can provide accuracy and simplicity simultaneously, are called Compact Models. Because usually the more accurate the model is, the more equations are required to describe device behavior, an accurate model takes longer computational time. Researchers invest great effort to balance the conflicting goals of model simplicity and accuracy. The research in this study is focused on MOSFET compact models, including statistical variability and reliability.

As stated in section 2.2, device variability has existed since the invention of ICs. Previously, process variability was dominant, and was included in SPICE analysis of circuits via 'corner' compact models and the worst-case approach. The devices that deviate from the nominal performance are used in the worst case. If circuits can perform well at the worst case, they should perform well in the normal situations. This method is very efficient when die-to-die variability is the leading variability, because to a degree this variability is deterministic. However, statistical variability is different from die-to-die variability. It is totally stochastic, as well as BTI, that trapped charges increase the statistical variability with time. The traditional worst-case analysis cannot capture the statistical property of circuit performance due to the stochastic nature of statistical variability and BTI. Therefore, large statistical compact models and circuit simulations are necessary in order to describe device and circuit behavior with statistical variability and BTI-induced ageing.

As the IC industry grew rapidly, SPICE software developed rapidly at the same time. A variety of compact MOSFET models appeared in different SPICE versions. These models were designed for the particular SPICE software and were not compatible with others, which was obviously an obstacle for compact model and SPICE development. In order to standardize compact models and help compact models to develop, the Compact Model Council was founded in 1995 [65]. Several of the compact models have been selected by the Compact Model Council as the standard models for SPICE and CMOS technology development. The most renowned one is the BSIM model [66] (Berkeley

Short-channel Insulated-gate field-effect-transistor Model), which is designed by the BSIM research group at UC Berkley [67]. It is widely implemented in a variety of SPICE simulators and has served the semiconductor industry for more than 20 years. BSIM4 belongs to the BSIM family, which started in the late 1980s, with the first one being the Berkeley Short-channel IGFET Model. Improvements were made through numerous iterations in the 20-year development of the BSIM family, with the motivation being "to develop a semi-empirical model which can cope with rapid changes and advancements in technology" [66]. BSIM4 is the extension of BSIM3, being more advanced in incorporating physical effects for sub-100 nm-regime MOSFETs. Another compact model, PSP, was attracting growing interest of the semiconductor industry. It is based on the surface potential and designed by Arizona State University and the NXP-TSMC Research Centre. Because BSIM4 is most intensively used compact model in industry, in this thesis, we focus on integrating the effect of statistical variability and BTI-induced ageing into the BSIM4 model.

2.5 SRAM

Static Random Access Memory (SRAM) plays a very important role in contemporary System-on-Chip (SoCs) [68]. SRAM usually occupies more than half of the die area and contains up to 2/3 of the transistors on a chip [69, 70]. It is at the top of the computer memory hierarchy because of its high speed, with write access times of less than 1 ns. Therefore, it is extensively used for caches or RAM in powerful microprocessors, micro-controllers, FPGAs, etc., all of which require burst transferring of the data. Transistor scaling is beneficial to SRAM, since aggressive scaling of transistor sizes enables maximum memory capacity on the smallest possible silicon area and directly reduces cost per chip. However, SRAM greatly suffers from statistical variability and BTI-induced ageing because its devices are minimally sized. Statistical variability and ageing makes each transistor intrinsically different. This greatly affects SRAM cell performance because of the matching requirements between the two inverters that form the core of the SRAM circuit. Moreover, large memories probe the statistical tails of transistor performance distributions, and the nature of these statistical tails governs bit failures when billions of devices are produced on one single chip. Therefore, SRAM circuit design requires large statistical-based compact models which include the influence of

statistical variability and BTI-induced ageing [71]. In this study, SRAM is used as a practical, industrially relevant vehicle to investigate the influence of statistical variability and ageing at circuit level.

Rather than charging or discharging capacitors, SRAM uses the positive feedback of back-to-back inverters to realize a bistable state. The standard 6-Transistor SRAM cell is shown in Fig.2.8. It includes two pass-gate (PG) transistors (PGL, PGR), two pull-up (PU) transistors (PUL, PUR) and two pull-down (PD) transistors (PDL, PDR). Except for the p-type pull-up transistors, all other transistors in the cell are n-type. The basic performances of the SRAM cell are 'read', 'write' and 'hold', shown as follows.

Hold: The binary state '1' or '0' is stored in the cell. In this situation, the wordline is low and pass-gates are disabled.

Read: During the read operation, bitline and not-bitline are pre-charged. When PG transistors are enabled, the sloping voltage of BL on the '0' side will be detected by a sense-amplifier attached to the bit lines, thus the stored data can be read out.

Write: While bit lines are kept to the state that is to be written, PG transistors open and the information is written into the cell, forcing the back-to-back inverters to one of their stable states.

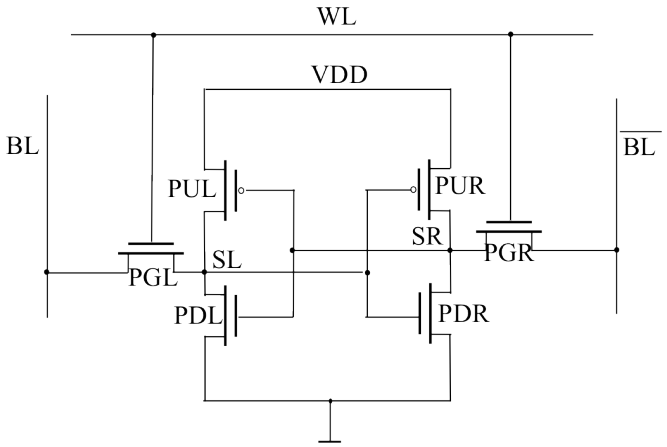


Fig.2.8 6T SRAM cell.

2.6 Summary

In this chapter, some most important challenges in MOSFETs scaling were introduced. As CMOS enters the sub-50 nm regime, statistical variability has proven to be key in device and circuit performance. The three main sources of statistical variability, RDD, LER and MGG have been discussed. Ageing effects due to BTI are considered, nothing that while BTI was initially observed in PMOS, the introduction of high-k metal gate technology has led to the observation of BTI effects in NMOS also. BTI is the critical reliability problem for MOSFETs and when it interacts with statistical variability, greater variation arises between devices. Since statistical variability and BTI-induced ageing greatly affect device performance, these effects should be introduced into compact models to help circuit design. Because they are inherently stochastic and make each device different, large statistic-based compact models and circuit simulations are required to simulate the circuit performance. In introducing SRAM it has been noted that it is very sensitive to statistical variability and BTI-induced ageing. The design of SRAM circuits increasingly requires compact models that integrate the effects of statistical variability and BTI-induced ageing, while also providing a mechanism to account for process variability.

Chapter 3

Research Methodology

3.1 Introduction

As device dimensions scale down aggressively, statistical variability becomes a dominant source of variations between transistors. It arises from the discreteness of charge and granularity of matter, making each transistor's electrical characteristic different from one to another [72]. Statistical variability on device performance is stochastic and thus the effect of it is stochastic as well. Meanwhile, HKMG technology exaggerates BTI. BTI greatly affects device reliability. Generation of defect states within the gate dielectric and subsequent trapping charges during device operation can lead to a significant increase in threshold voltage, lowering the drive current. In addition, the dielectric layer becomes easier to break down. Moreover, the interplay between statistical variability and BTI results in further device parameter variation which stochastically affects the degradation in performance as a function of device ageing. The degradation in device performance subsequently degrades the circuit performance, increases the power dissipation and decreases the yield. The effects of statistical variability and BTI-induced ageing were introduced in detail in Chapter 2. Integrating these effects into compact models is necessary. In particular, this is important for the Design Technology Co-Optimization (DTCO) and the corresponding tool flow and empowers variability aware circuit design. A statistical compact modelling methodology that accurately captures the statistical properties of variability and BTI degradation is a prerequisite to guarantee the effectiveness of circuit level evaluations.

The accurate modelling of the impact of device variability and reliability on circuit behavior requires compact models to meet the following standards. Firstly, compact models should capture the distributions of device figures of merit (the performance index of each device, introduced in section 3.4.2). This indicates that the compact models are capable of modelling the behavior of variable devices at different ageing levels. Secondly, compact models should capture the correlations between device figures of merit and hence the behavior of the technology is preserved in the compact model extraction strategy. Thirdly, as large-scale circuits contain billions of transistors, compact models that meet the first two goals should be generated sufficiently large in order to provide detailed information in the distribution tails following statistical circuit simulation and extraction of statistical circuit figures of merit therefrom [73]. Failure to have a sufficiently large independent sample of compact models for circuit simulation will result in subsampling and generation of statistical artefacts in output distributions. This will be introduced in section 5.2 in detail. Only when compact models meet these three criteria, statistical circuit analysis can be meaningful and power/performance/yield (PPY) can be accurately predicted.

In order to accommodate these criteria in compact model extraction and generation, it is important to develop systematic simulation methodologies to enable the statistical circuit simulation. The description of the simulation methodologies and tools is the main subject of this chapter.

Section 3.2 describes the research flow and simulation tool chain used in this study to facilitate each step. The following sections outline the methodologies used in each step, except for the compact model generation methodology that will be introduced in detail in Chapter 5. In section 3.3, the physical simulation methodology, based on Drift Diffusion (DD) approach, and used to obtain device performance under statistical variability and ageing conditions, is presented. Section 3.4 discusses the compact model extraction methodology. The statistical circuit simulation engine RandomSpice, that is employed in this study to perform statistical circuit simulation, is explained in section 3.5. Section 3.6 summarises this chapter.

3.2 Research Flow and Simulation Tool Chain.

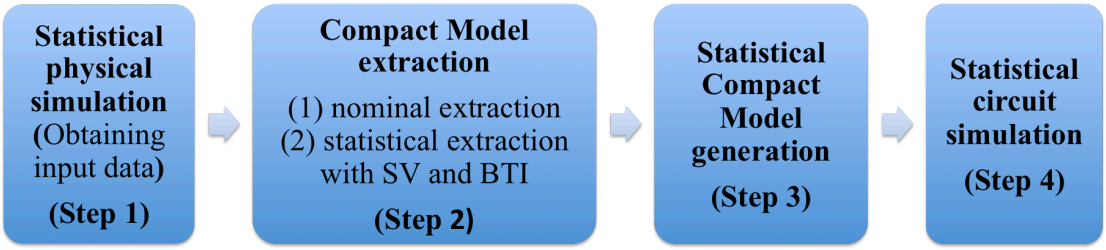


Fig.3.1 The simulation flow.

The simulation flow of this research is shown schematically in Fig.3.1, which is aimed at generating accurate compact models that are capable of describing the effects of statistical variability and BTI-induced ageing and assessing these effects on circuit performance. The initial step is the statistical physical transistor simulation (step 1), followed by statistical compact model extraction (step 2), statistical compact model generation (step 3) and finally, statistical circuit simulation (step 4). In order to extract and generate compact models, transistor characteristics under the influence of statistical variability and BTI-induced ageing are required. Therefore, obtaining this data as an input for compact model extraction is the first step, with physical TCAD simulations being used for this purpose. Then statistical compact model extraction follows in step 2. The number of extracted compact models is limited to the simulation ensemble size from step 1. The distribution of the extracted compact model parameters are analysed and used to generate sufficiently large compact models on the fly in step 3. The limited number of extracted compact models is not directly used because this results in subsampling within statistical circuit simulation, details of which will be introduced in section 5.2. With the methodology used in step 3, generated compact models will preserve both the distribution of the SPICE model parameters from step 2 as well as the correlation between them. The compact model generation methodology is a feature implemented within the Monte Carlo circuit simulation engine RandomSpice. This methodology gives us the ability to generate a large number of compact models, sufficiently large to simulate circuits of a size typical to modern commercial circuit simulations (for instance of the order of 10^5 components). The methodology is designed so that this large number of compact models are sufficiently accurate to give accurate circuit simulation results. Then we apply them to statistical circuit simulation. Here, models are generated to

investigate the influence of statistical variability and BTI-induced ageing on SRAM cells (step 4). SRAM cells are chosen for this study as their design and margining are critical in modern microprocessors since SRAM occupies a large fraction of the chip area and individual cells are highly sensitive to variation.

The TCAD-based Design Technology Co-Optimization (DTCO) GSS tool chain is used in this study to facilitate this research, as shown in Fig.3.2. This tool chain is composed of four main software applications. These are the structure translator Monolith, the 3D statistical TCAD simulator GARAND, the statistical compact model extractor Mystic and the statistical circuit simulator RandomSpice [74]. For this study, a well-designed 25 physical gate length MOSFET has been used, the details of which will be introduced in section 4.2.1. In addition to the TCAD simulation tools, this work makes use of powerful computing clusters with a database for management of the large volume of data generated by statistical device simulations. These clusters includes 1016 cores featuring dual Intel Xeon E5530 processors with 48 GB RAM per node and 2376 cores ranging from dual Intel Xeon X5650 processors with 48 GB RAM per node to dual Intel Xeon E5-2650 v2 processors with 64 GB RAM per node. Large sample of statistical device simulations, compact model extractions and circuit simulations can be submitted to the cluster via the advanced job management system. Jobs can be executed simultaneously and efficiently using this large number of processors.

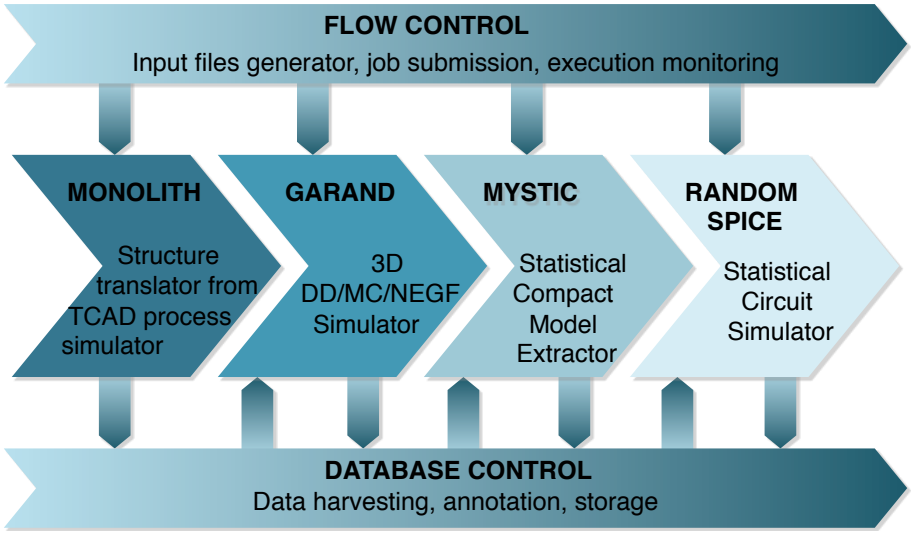


Fig.3.2 The TCAD based Design Technology Co-Optimization (DTCO) GSS tool chain.

3.3 Physical Simulation Methodology

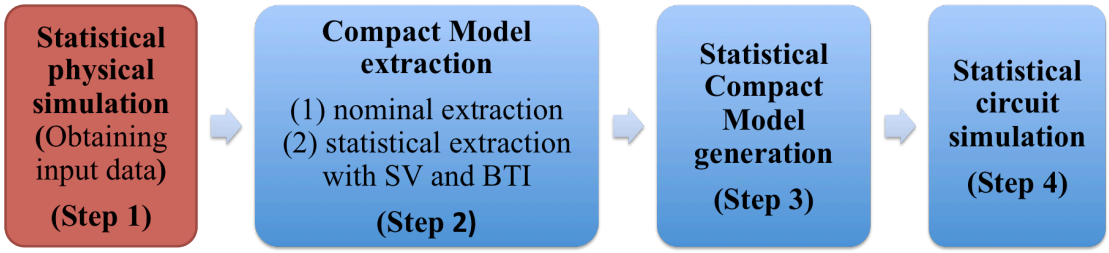


Fig.3.3 Step 1 of the simulation flow.

In order to generate statistical compact models, a large ensemble of current-voltage (IV) characteristics, which describe the individual device performances under the influence of statistical variability and BTI-induced ageing, are required (Step 1 of the simulation flow shown in red in Fig.3.3). Individual compact models are simply extracted against this data, with one model produced for each device in the initial physically simulated ensemble.

Two approaches can be used to obtain the statistical transistor IV characteristics. The first approach is a direct measurement of devices from test chips. However, this method is limited to the existing generation of transistors and is time and resource consuming. It also may not be applicable to others when the transistor characteristics are rapidly changing with ageing. The second approach is to obtain data by means of physical simulation. This approach is capable of providing data for both existing and future device technology generations. The disadvantage of this method is that the data is based on the best available process and transport models, which may result in differences compared to the real device performance. These differences can be reduced by careful calibration to experimental data. The common way for industry is to use the physical simulation to predict and optimize device design at the early stage of new technology development. When the new technology becomes available, calibration is performed using data measured from real fabricated devices to improve the physical models.

With regard to this research, not only fresh devices (containing only statistical variability), but also devices at different BTI-induced ageing levels are required. It is

difficult to age devices to the exactly expected ageing levels, as well as to the end of its life conditions. Physical simulation can give a clear indication for the cause of any change of the electrical characteristics. For example, device performance at the trap density of $5x10^{11}cm^{-2}$ can be easily obtained by physical simulation. Meanwhile, the 20 nm CMOS technology used as an example in this research was still in conception when this research started. Therefore, physical device simulation is used in this study.

Physical device simulation is based on fundamental device physics, deriving electrical characteristics from device structure. There are mainly three approaches to physical device simulation. These are, in order of increasing physical complexity, Drift-Diffusion (DD) [75, 76], Monte Carlo (MC) [77] and Non-Equilibrium Green's Function (NEGF) [78]. There are three main components of the DD model for semiconductor device simulation: Poisson's equation that solves the potential distribution (electrostatic) within the device, including the carrier density statistics that relate the electron and hole densities to the local potential, and current continuity equations in drift-diffusion approximation with the local mobility model that defines the carrier transport. The main advantage of the DD approach is that it is computationally inexpensive compared to the other models. MC simulation is a semi-classical approach that directly solves the Boltzmann transport equation (BTE) by simulating the ballistic motion of carriers within the device, interrupted by stochastic scattering events. Scattering rates are based upon quantum mechanical transition rates based typically upon Fermi's Golden Rule. The advantage of MC over DD is in the intrinsic ability to resolve non-local and non-equilibrium transport, which is critical for accurate simulation of modern ultra-small devices. However, it is very time consuming and as it is a statistical method, MC is unsuited to simulate in the sub-threshold regime where currents are very small. NEGF is a full quantum mechanical transport solution and gives accurate information on tunnelling and leakage, which is of importance in minimising power, but is the most computationally demanding approach and often does not include scattering.

In this research, the statistical 3D 'atomistic' simulator GARAND is employed to perform physical simulations. GARAND is designed for predicting present and future CMOS transistor performance in the presence of statistical variability and reliability. DD, MC and NEGF modules are all included in GARAND.

Since BTI-induced ageing is primarily an electrostatic effect and well captured by the DD model, DD is a good choice. In addition, device simulation is no longer restricted to a single device performance because each transistor's performance varies due to statistical variability and accumulation of random traps. Simulations of thousands of devices have to be performed at different ageing levels in order to obtain detailed information on the extreme ends of the distribution of device figures of merit. Due to the fact that statistical variability and BTI-induced ageing are intrinsically 3D in nature, 3D device simulations of large ensembles are necessary to fully capture the influence of carrier trapping on devices. This indicates a requirement for large computational power and time. Therefore, employing quantum corrected DD module is the best choice of the simulation technique in this research. To take into account the effects of statistical variability, the major sources of variability, including RDD, LER, MGG and BTI-induced ageing, are introduced in the GARAND simulator. The following sections present the details of the DD method and the simulation methodologies for statistical variability and BTI-induced ageing.

3.3.1 Drift Diffusion

The current continuity equation in DD approximation is given by equation (3.1):

$$\nabla . J_n = 0 \tag{3.1}$$

where J_n is the current density vector.

In the DD model, the sum of drift current (induced by the electric field) and diffusion current (resulted by the carrier density gradient) approximates the total current density of electrons and holes, as given by (3.3) and (3.4).

$$J_n = qD_n \nabla n - q\mu_n n \nabla \psi \tag{3.2}$$

$$J_p = -qD_p \nabla p - q\mu_p p \nabla \psi \tag{3.3}$$

Where J_n represents n-channel MOSFET current and J_p represents p-channel MOSFET current. D is the diffusion coefficient and μ is the corresponding carrier mobility. ψ is the electrostatic potential, n and p are the electron and hole carrier densities.

The potential in equations 3.2 and 3.3 is obtained by solving the Poisson equation:

$$\nabla \cdot (\epsilon \nabla \psi) = q(n - p + N_A^- - N_D^+) \quad (3.4)$$

where ϵ is the intrinsic permittivity. N_A^- and N_D^+ are the ionised acceptor and donor doping concentrations respectively. In n-type device the majority carriers are electrons and in p-type device the majority carriers are holes.

The DD approach benefits from its low computational cost. However, the decananometer scale MOSFETs greatly suffer from quantum effects and the inherent simplifications of DD technique limit the accuracy of this approach. Therefore, Density Gradient (DG) quantum corrections are applied in GARAND to take the quantum effects into account that will be described in section 3.3.2 [22].

3.3.2 Density Gradient Corrections

In order to improve the accuracy of device simulations in the deep submicron regime, DG corrections are introduced in the drift-diffusion module of GARAND to capture the significant impact of non-local quantum effects [79]. Important quantum mechanical phenomena, including the quantum confinement effects that dominate the latest device architectures, can be accurately described with DG corrections [79-81]. The corrections are introduced by quantum potential ψ_{qm}, which is proportional to the second derivative of the carrier density as shown in Equation (3.5).

$$\psi_{qm} = 2b_n \frac{\nabla^2 \sqrt{n}}{\sqrt{n}} = \phi_n - \phi + \frac{k_B T}{q} \ln\left(\frac{n}{n_i}\right) \quad (3.5)$$

Where b_n is the expression of the density gradient dependence. The Poisson equation with the DG corrections is shown in Equation (3.6).

$$J_n = qD_n \nabla n - q\mu_n n \left(\nabla \psi + 2\nabla \left(b_n \frac{\nabla^2 \sqrt{n}}{\sqrt{n}} \right) \right) \quad (3.6)$$

The Gummel decoupled method is applied in GARAND for the solution of the above equations [including Poisson equation (3.4), DG corrections (3.5), and current continuity equation (3.1)]. Equations are solved in succession and the result obtained by solving one equation is used in the next equation for the seeking of the new solution. The iterations are performed until the system converges. In this process, Poisson equation and DG correction equation are solved using a Newton Successive Over Relaxation (SOR) solver. The current continuity equation is solved using a Bi-Conjugate Gradient Stabilized

(BicGSTAB) solver. With the approach above, the 25 nm gate length bulk silicon devices used in this research are accurately simulated for IV characteristics [20, 22].

3.3.3 Incorporation of Statistical Variability and Ageing

The effects of three main sources of statistical variability (RDD, LER, and MGG) and BTI-induced ageing greatly affect planar MOSFET performance, as described in detail in section 2.2 and 2.3. This section describes the methodologies of incorporating these effects into the 3D simulator 'GARAND'.

RDD has been introduced in GARAND by employing the methodology as described in [82]. The placement of a random dopant at a silicon lattice site is determined by the local ratio of the doping and silicon atom concentration at that site. A Cloud-in-cell (CIC) scheme [83, 84] is used to spread the dopant charge to the surrounding eight mesh points of the cube. The amount of the charge to a particular discretisation node is determined by the distance between the dopant and the node. The charge assigned to a particular mesh point $p(x_p, y_p, z_p)$ is calculated using the equation below:

$$\rho(x_p, y_p, z_p) = w_{x_p} w_{y_p} w_{z_p} \frac{1}{v} \tag{3.7}$$

where ρ is the charge assigned at that mesh point. v is the volume of the cell given by $v = h_x h_y h_z$ (h_x, h_y and h_z are the mesh step size on x, y, and z axis respectively). $w_{x_p} w_{y_p} w_{z_p}$ are the weight factors depending on the distance between the charge and the mesh point on x, y and z axis respectively. The equation to calculate the weight factor is $w_{x_p} = h_x - |x - x_p|$ if $|x - x_p| \leq h_x$. If $|x - x_p| > h_x$, then the weight factor $w_{x_p} = 0$. The potential distribution affected by RDD in bulk MOSFET is shown in Fig.3.4.

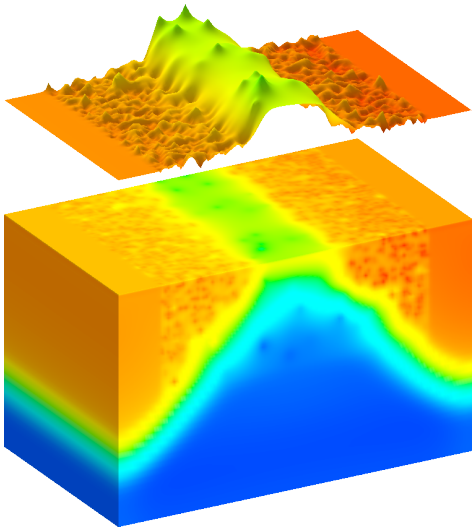


Fig.3.4 The potential distribution of bulk MOSFET with RDD. Used with permission.

Line Edge Roughness (LER) is introduced in GARAND by using 1D Fourier synthesis. The gate edge is generated by the power spectrum associated with the Gaussian autocorrelation function [50]. The Gaussian power spectrum equation is given by Equation (3.8).

$$S_G(k) = \sqrt{\pi}\Delta^2 \wedge exp\left(-\frac{k^2\Lambda^2}{4}\right) \tag{3.8}$$

Where $k = i\left(\frac{2\pi}{Ndx}\right)$, $0 \leq i \leq \frac{N}{2}$. N is the number of the mesh points and dx is the mesh step size in Fourier space. The Root Mean Square (RMS or Δ) and correlation length (Λ) are two key parameters. RMS indicates the average vertical deviation of the line roughness and correlation length (Λ) represents the distance over which the fluctuations are correlated. The autocorrelation function is the inverse Fourier transform of the power spectral density function. The gate edges are generated from the power spectrum corresponding to the Gaussian auto correlation function. The RMS amplitude of 4 nm with a correlation length of 30 nm is applied in this study. The potential distribution of bulk MOSFET under the influence of LER is shown in Fig.3.5.

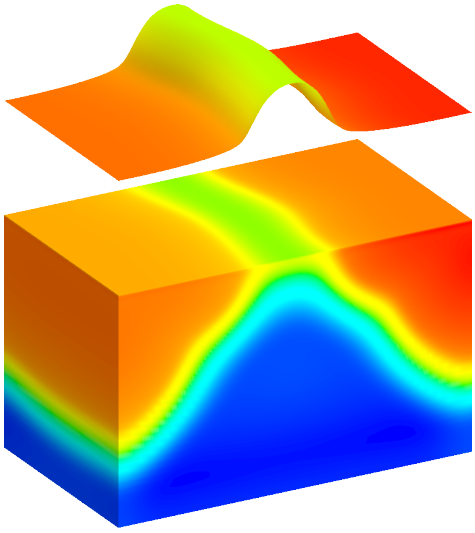


Fig.3.5 The potential distribution of bulk MOSFET with LER. Used with permission.

MGG is introduced by generating granular structure of the metal gate by Voronoi Tessellation technique [85]. The Voronoi Tessellation is the method that can partition the plane into different size of regions based on the distance to the seeds. The distance from every point of the region to the seed in that particular region is smaller than any other seeds. In this research, the average grain diameter is 7 nm. Different grains are assigned to different work functions randomly. The potential distribution of bulk MOSFET under the influence of MGG is shown in Fig.3.6.

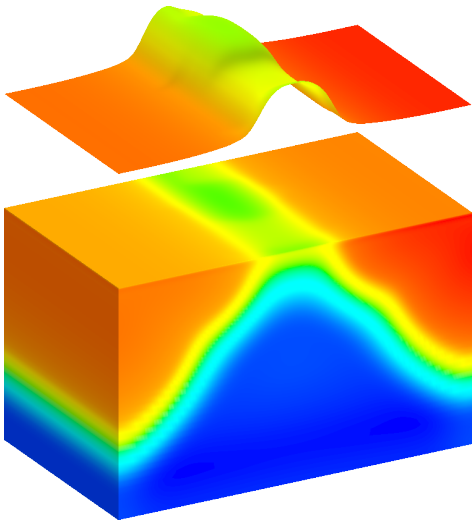


Fig.3.6 The potential distribution of bulk MOSFET with MGG. Used with permission.

The effects of ageing as a result of BTI-induced traps is introduced into GARAND by imposing a fine auxiliary 2D mesh at the interface of Si and gate dielectric. Similar to RDD, at each node of this mesh, a rejection technique is used to determine the presence of a single charge depending on the charge density. Then the cloud-in-cell scheme is applied to the charge to assign its charge to the surrounding nodes of the mesh [60].

The incorporation of statistical variability and BTI-induced ageing into GARAND enables the effects associated with them to be included in statistical compact models by providing target data for models to be extracted against. The simulations performed by GARAND provide this data for compact model extraction in this work.

3.4 Compact Model Extraction Methodology

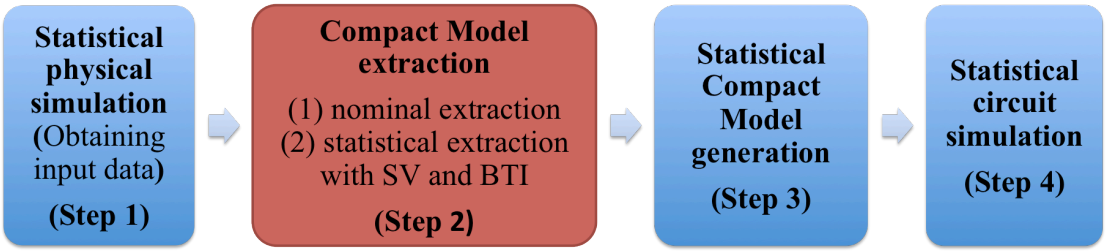


Fig.3.7 Step 2 of the simulation flow.

This section discusses the methodology that allows the accurate compact models to be extracted based on the physical simulation results (step 2 of the simulation flow, shown in red in Fig.3.7). It is the bridge that transfers the IV characteristics obtained from physical 3D simulations (step 1) to the compact models that can be used in circuit simulations. It is expected that with this methodology, extracted compact models could reproduce the transistor figures of merit and the correlations between them from the physical simulation. Meanwhile, some information can be extracted by analysing the extracted compact model parameters for step 3 of statistical compact model generation. The compact model extraction is facilitated by the GSS software Mystic [86], which provides multiple optimisers and allows flexible extraction strategies.

In this research, the industrial standard model BSIM4 developed by UC Berkeley [2] (BSIM4 is detailed in section 2.4), is selected as the compact model. BSIM4 is a threshold voltage based semi-empirical model with a wide range of parameters and semiconductor equations to represent the device behaviours. Some of the parameters are with physical meaning, while others are non-physical fitting parameters. The use of physical parameters through the whole extraction strategy leads to an estimation of the extracted parameters within a reasonable domain. Non-physical parameters are used to fit phenomenologically the IV characteristics. This retains the physical meaning on the physical parameters. This is vitally important for compact model generation, as compact model generation is based on generating a careful selection of physical parameters that could capture the influence of statistical variability and BTI.

Statistical compact model extraction is implemented by two stages, as originally described in [87]. The first stage is the uniform model extraction with the fixed and exact length, width, and continuous doping profile, when no statistical variability is considered. The second stage is the statistical compact model extraction, in which statistical variability and ageing effects are injected into the uniform compact model from stage one. As this research is based on the 25 nm gate length bulk MOSFET, the 25 nm statistical BSIM4 compact models with the effects of statistical variability and ageing is the target of this section. The extraction of BSIM4 model cards are based on a full set of electrical characteristics obtained from simulations performed by GARAND (or can be obtained by experimental data).

3.4.1 Uniform Model Extraction

The uniform compact model is based on the ideal device, excluding statistical variability on geometry and continuous doping profile. At this stage, ageing information is not taken into account. It is vitally important to obtain an accurate uniform compact model, since it greatly influences the precision of the statistical compact model and ultimately the accuracy of the statistical circuit simulations. It is expected that parameters with physical meanings can be retained as much as possible while the non-physical parameters are sacrificed and changed in the whole process in order to obtain the best fit.

There are two extraction approaches: group extraction and single device extraction. Group extraction approach aims at extracting a group of devices with different geometries. Single device extraction targets at extraction of a specific device geometry performance. Group extraction approach is more advanced in capturing parameters' physical meaning while single device extraction is aimed at fitting the phenomenological curve. Comparing with single device extraction, group extraction may not be perfect for any single device, but it offers more accurate results while fitting a group of devices with different geometries, since it remains parameters' physical meaning. Therefore, the group extraction approach is applied in this research.

Before uniform model extraction starts, physical simulations are performed by GARAND at multiple channel lengths, device widths, different substrate bias, high drain (1V) and low drain (0.05V) bias. The temperature in all simulations or extraction is at 27°C. Table 3.1 shows all structured parameters related to all device geometries and bias conditions. The basic transistor IV characteristics, including I_D-V_G at low and high V_{DS} bias and different V_B, I_D-V_D at different V_{GS} bias and V_B=0 are needed in the compact model extraction process.

Although the target of the uniform model is the 25 nm gate length bulk MOSFET, the initial extraction is done on long channel devices in order to obtain parameters that are independent of short channel effects [88], such as VTH0 (low drain threshold voltage), VOFF, RDSW (source/drain resistance), UA (mobility) and MINV (middle inversion). Short channel devices are used to extract parameters related with short channel effects, for example ETA0 (DIBL). Different channel lengths and widths are considered in this extraction in order to make the extraction approach scalable.

Table 3.1 The physical simulation scenarios for uniform compact model extraction.

Channel length	200 nm, 150 nm, 100 nm, 50 nm, 40 nm, 30 nm, 25 nm, 20 nm
Device width	30 nm, 25 nm, 20 nm
Drain bias	1.00V, 0.05V
Body bias	-1V, -0.8V, -0.6V, -0.4V, -0.2V, 0V
Temperature	27°C

To start the extraction, some initial process parameters containing the device information are required. These parameters include dielectric constant and gate oxide thickness (TOXE), doping concentration in the channel (NDEP), temperature (TNOM), mask level channel length (L_{drawn}), mask level channel width (W_{drawn}), and source-drain junction depth (XJ). The values for the 25 nm device are shown in Table 3.2.

Table 3.2 Initial parameter values before extraction for 25 nm device.

Parameter	NMOS	PMOS
TOXE	$8.5x10^{-10}$	$8.5x10^{-10}$
NDEP	$6.32x10^{18}$	$7.047x10^{18}$
TNOM	27	27
L_{drawn}	$25x10^{-9}$	$25x10^{-9}$
W_{drawn}	$25x10^{-9}$	$25x10^{-9}$
XJ	$1.75x10^{-8}$	$1.5x10^{-8}$

The optimization method used for the compact model extraction is the combination of trust region iteration and linear-squares fit. During the extraction process, local optimizing is performed at each parameter extraction, and the process is repeated until the best fit, determined by the lowest error, is achieved. This method enables to retain the physical meaning of the extracted parameters since each parameter is independently extracted within the physical meaning domain. Global optimization might result in the best fit of the IV characteristics, however, this leads to parameter compensation due to high correlations and complex interdependence between the extracted parameters. This will lead to problems in the compact model generation, which relies on parameter distributions that are expected to be continuous and mono-modal. Global optimization

cannot fulfil our main aim to keep the physical parameters as original as possible and change the non-physical parameters to help fit. Thus, global optimization is not considered in this research. Fig.3.8 and Fig.3.9 present the BSIM4 results of the 25 nm gate length n-MOSFET against the GARAND simulation results.

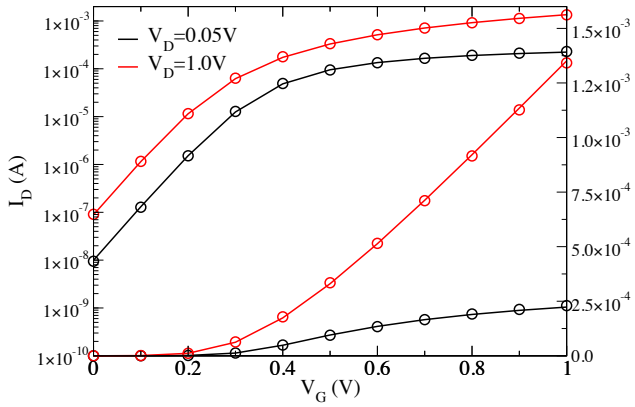


Fig.3.8 I_D-V_G characteristics of the 25 nm BSIM4 nMOSFET model. Solid line: simulation results by GARAND. Symbol: the extracted compact model values.

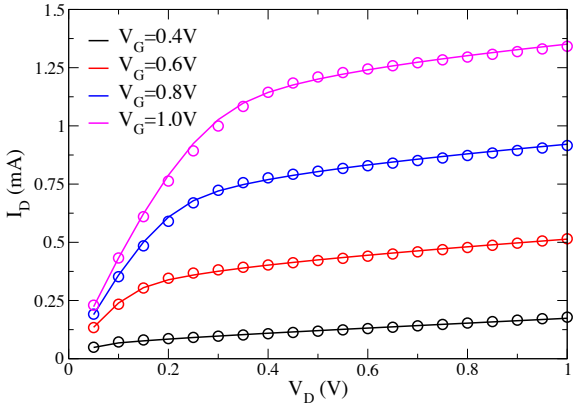


Fig.3.9 I_D-V_D characteristics of the 25 nm BSIM4 nMOSFET model. Solid line: simulation results by GARAND. Symbol: the extracted compact model values.

The extracted models include accurate substrate bias dependence at both low and high drain bias, illustrated in Fig.3.10 and Fig.3.11, where the device behaviour is well maintained for substrate biases of 0, -0.2, -0.4, -0.6, -0.8 and -1.0V.

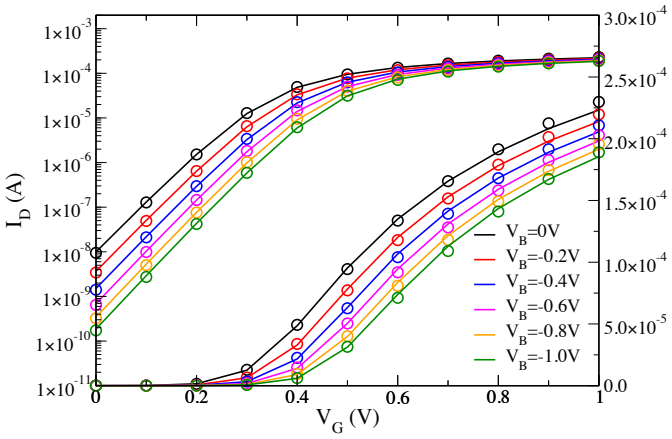


Fig.3.10 BSIM4 results of 25 nm nMOSFET at V_{DS} = 0.05V for substrate biases of 0, -0.2, -0.4, -0.6, -0.8 and -1.0V. Solid line: simulation results by GARAND. Symbol: the extracted compact model values.

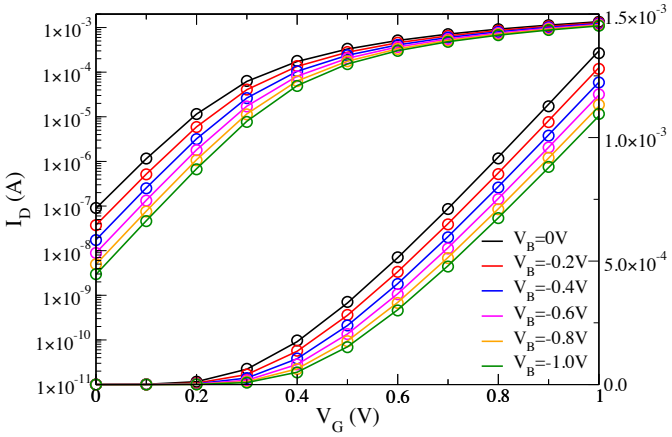


Fig.3.11 BSIM4 results of 25 nm nMOSFET at V_{DS} =1.0V for substrate biases of 0, -0.2, -0.4, -0.6, -0.8 and -1.0V. Solid line: simulation results by GARAND. Symbol: the extracted compact model values.

The extracted 25 nm uniform model, that can accurately capture the body bias and drain bias dependence, follows the approach described in [89, 90] and is directly used in this study. In order to include statistical device variation and ageing in circuit simulation, we need to go beyond the above compact model extraction and extract statistical models against each device within the simulated statistical device ensemble. This is the second stage of compact model extraction and requires as a starting point the above uniform compact model, which was the first stage. The second stage, statistical compact model extraction will be discussed in detail next.

3.4.2 Figure of Merit Based Statistical Compact Model Extraction

The figures of merit define important aspects of the transistor operation. Table 3.3 shows the definitions of the key figures of merit for a MOS transistors. The figures of merit capture the most important characteristics of the device. It is convenient to analyse a particular aspect of the transistor operation based on the representation the particular feature figure of merit represents. The key figures of merit listed in Table 3.3 are the most important figures in respect to MOSFETs, including the threshold voltage, drive current, leakage current and short channel effect respectively. These parameters play an important role in circuit design and analysis.

Table 3.3 The definitions of figures of merit.

V_{TH} at high drain bias	Gate voltage when drain current reaches a criteria at high drain bias. (In this study, $1\times10^{-7}A/\mu m$ is used as the criteria and the high bias in this research is 1V)
V_{TH} at low drain bias	Gate voltage when drain current reaches a criteria at low drain bias. (In this study, $1\times10^{-7}A/\mu m$ is used as the criteria and the low bias in this research is 0.05V)
I_{ON} at high drain bias	Drain current when gate and drain are at the high bias.
I_{ON} at low drain bias	Drain current when gate is at high bias and drain is at low bias.
I_{OFF} at high drain bias	Drain current at high drain bias when gate voltage is 0.
I_{OFF} at low drain bias	Drain current at low drain bias when gate voltage is 0.
DIBL	The subtraction of high drain and low drain's V_{TH}

The statistical compact model extraction is aimed at producing compact models that represent device behaviours under statistical variability and BTI-induced ageing, while also capturing the device figures of merit as well as the correlations between figures of merit within the device ensemble. In this work, statistical device simulations using

GARAND are performed in order to provide the target device data for the extraction of statistical compact models.

The second stage extraction flow is shown in Fig.3.12. Parameters that are capable of capturing statistical variability and ageing effects are selected. These parameters can shift the corresponding compact model's performance to fit the changed device performance due to statistical variability and ageing. The selected parameters are re-extracted from each device simulated by GARAND considering statistical variability and different levels of BTI-induced ageing. Since device performance varies due to the influence of statistical variability and ageing, each set of re-extracted parameters for each device varies from each other. These sets of parameter values are inserted into the uniform model extracted at stage one to substitute the same parameters' values in the uniform model. In other words, only the re-extracted parameters are replaced in the uniform model, other parameters remain the same as in stage one. Each set of the re-extracted parameters from one simulated device produces one compact model. Thus, compact models are successfully extracted and form the lookup table models.

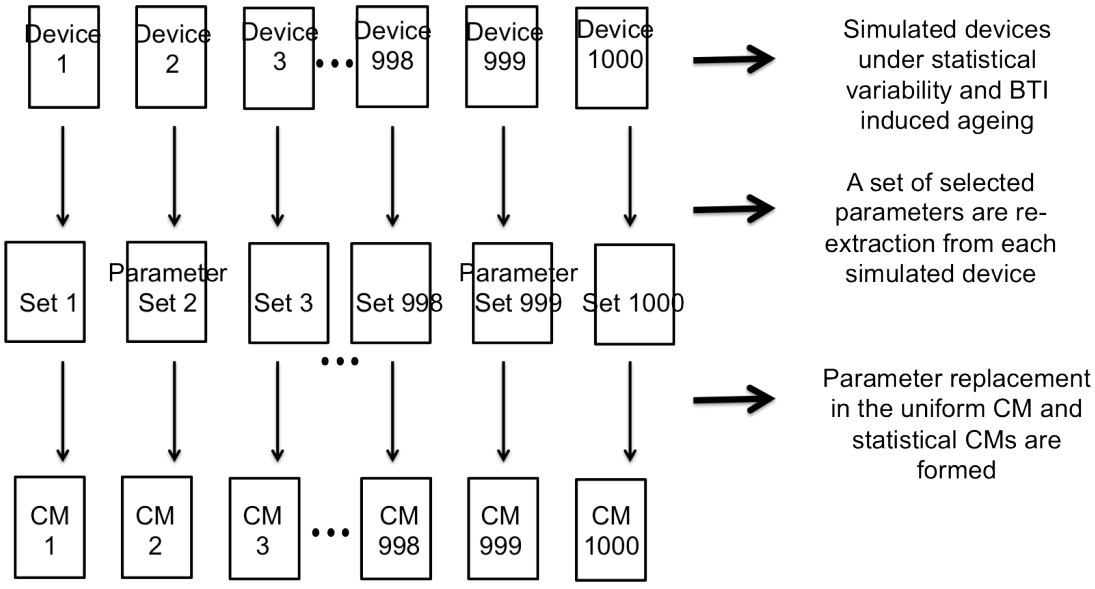


Fig.3.12 Statistical compact model extraction (stage-two) flow.

Since every model is formed by inserting several parameter values re-extracted from the simulated devices, the extracted models are limited by the number of devices in the

physical simulations. The limited number of compact models will result in subsampling in statistical circuit simulation (introduced in section 5.1). Therefore, very large ensembles of compact models (up to 6-sigma depending on the application) are required to be generated and used in the circuit simulation. “Model generating” means that new compact models are constructed rather than directly extracted from physical simulation. Since selected parameters are re-extracted and replace the same parameter value in the uniform model in stage two, it is reasonable to think of generating the new values that could follow the distribution of each re-extracted parameter and also present their correlations. Then, these newly generated values will be inserted into the uniform compact model in the same way as in stage two. Thus, the number of compact models that can be generated is unlimited.

The compact model generation method emphasises the importance of the parameter selection at stage two. Essentially, selected parameters need to meet two requirements: 1) They can accurately present the influence of statistical variability and ageing for each device. 2) The number of selected parameters should be as small as possible for the simplicity of compact model extraction and generation. Since the influence of statistical variability and ageing on device performance is captured by the device figures of merit, parameters can be selected in the way that each parameter targets each individual figure of merit. From the previous research [87, 89-91] and with a deep understanding of BSIM4 parameters, seven parameters are selected for this study. They should be sufficient to represent the statistical variability and ageing effects on the transistor behaviours in this research. The parameters are VTH0, EAT0, VOFF, NFACTOR, UA, VSAT and CDSCD. The physical meanings behind these parameters are listed in Table 3.4.

Table 3.4 Descriptions of the compact model parameters.

Selected Parameter	Description of the selected parameter
VTH0	Low drain threshold voltage
ETA0	DIBL
VOFF	Off current at low drain
NFACTOR	Low drain subthreshold slope
UA	Mobility
VSAT	On current at high drain
CDSCD	High drain subthreshold slope

The extraction of the above parameters is performed against the I_DV_G curve of each simulated device under statistical variability and BTI-induced ageing. These selected parameters should capture the differences between uniform devices and devices under statistical variability and BTI-induced ageing. The selected parameter extraction flow is shown in Fig.3.13. VTH0 is the first re-extracted parameter. It can linearly shift threshold voltage in compact model at low drain voltage and therefore is used for capturing the change of V_{TH} resulting from statistical variability and BTI-induced ageing at low drain voltage. ETA0 targets at DIBL. Therefore, when VTH0 is extracted, ETA0 can adjust the threshold voltage at high drain bias performance. VOFF captures the off-current and NFACTOR captures the subthreshold slope at low drain voltage. These two factors influence each other and are usually extracted together to ensure the accuracy of both parameters. UA is extracted to capture the variation in carrier mobility. VSAT calibrates the on current at high drain. CDSCD is the last extracted parameter and is used to adjust the subthreshold slope at high drain voltage.

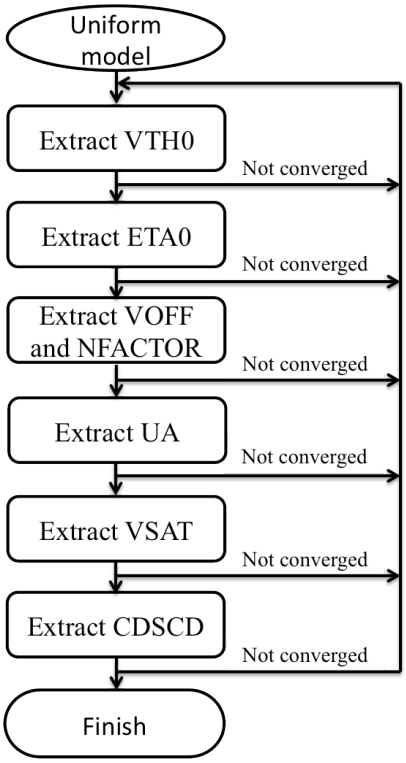


Fig.3.13 Selected parameter extraction flow at stage-two.

The results of this approach are presented in detail in Chapter 4. The accuracy of this approach is verified by comparing the extracted compact models against the physical simulation. This approach successfully transfers statistical variability and BTI-induced ageing from physical simulations to compact models, which will be presented in Chapter 4.

Compact model generation methodology is based on the analysis of the compact model extraction results, especially the parameters extracted at the second stage. Therefore, the compact model generation methodology (at the third step of the research flow) will be introduced in detail in Chapter 5 after analysing the extraction results in Chapter 4.

3.5 Statistical Circuit Simulation

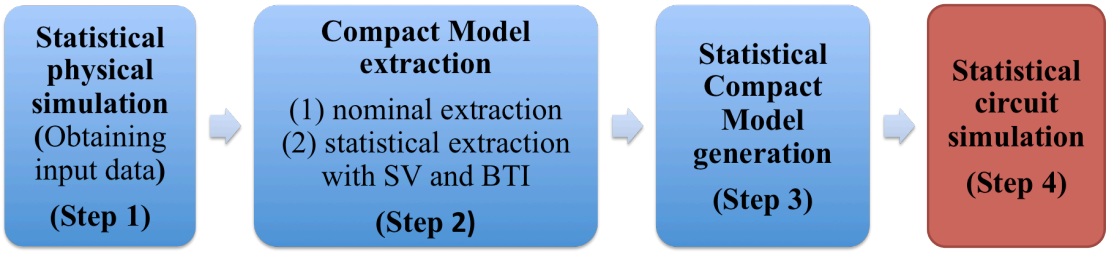


Fig.3.14 Step 4 of the simulation flow.

Statistical circuit simulation is the last step of the simulation flow (shown in Fig.3.14). As the interplay of statistical variability and BTI-induced ageing effect becomes increasingly critical in 65 nm technologies and beyond, performance of identified devices in a circuit varies [92], stochastically. As a result, a critical reduction of yield (defined as the proportion of the manufactured chips that function correctly) may occur in a traditional design [93]. The traditional method used in the circuit design is the worst-case analysis, which has been introduced in section 2.4. This method results in pessimistic circuit predictions. Because of the random nature of device performance due to statistical variability and BTI-induced ageing, the better way in the contemporary circuit design is to perform Monte Carlo (MC) circuit simulation. Large statistical circuit simulations are necessary to help predict, analyse and optimize design performance and yield under the influence of these effects, especially to investigate the circuit behaviour far into the extreme tails of the circuit characteristics' distribution.

Therefore, the MC circuit simulation engine, RandomSpice (the last software in the simulation tool chain) is used in this study. It supports a wide range of SPICE simulators to perform MC circuit simulation. The open source ngSPICE is used in this study. MC method is adopted to capture the distribution to the population by simulating a large number of circuits. MC method obtains circuit performance distribution by repeating simulations of random sample circuit [94]. Thus, in the large-scale MC circuit simulation, by repeating random sampling devices with different characteristic, circuit performance and yield can be well predicted. The more repetition the simulation performs, the more accurate the result will be. Thus, a large device sample is required to

repeatedly perform circuit simulation with random sampling devices with different characteristics. Therefore, sufficiently large compact models are necessary for RandomSpice to randomly sampling. Facilitated by high performance computing clusters, large-scale circuit simulations can be simultaneously performed with RandomSpice. This allows circuit designers to investigate rare transistor performance and its influence at circuit level. It also gives a good yield prediction of circuits containing billions of transistors.

Making a compact model library is the first step of the circuit simulation using RandomSpice. Different keywords have to be specified to create different types of compact models when making the compact model library. These keywords appear after the MOSFETs, which need to be randomly sampled in SPICE netlist, telling RandomSpice to randomly generate the corresponding compact models using information allude in the library. In RandomSpice, several compact model generation methods are embedded, including Gaussian V_T, Principal Component Analysis (PCA), etc. The new generation method used in this study (detailed in Chapter 5) is embedded in RandomSpice as well for the investigation of the effects of statistical variability and BTI at circuit level. Statistical circuit simulation is performed using SRAM as the vehicle and the results are analysed in Chapter 6.

3.6 Summary

In this chapter, the simulation flow and methodologies adopted to perform each step of the flow are introduced. The input data used to extract compact models are obtained by physical simulations using DD module with DG quantum corrections in GARAND device simulator. The two-stage statistical compact model extraction methodology is introduced in section 3.4. This is the second step of this research and is facilitated by the statistical compact model extractor in the simulation tool chain. The statistical circuit simulations enable investigation of the influence of statistical variability and BTI-induced ageing by embedding compact model generation methodology (presented in Chapter 5) in RandomSpice. In the following chapters, we will report the simulation results obtained by means of the methodologies introduced in this chapter.

Chapter 4

Physical Simulation and Compact Model Extraction

4.1 Introduction

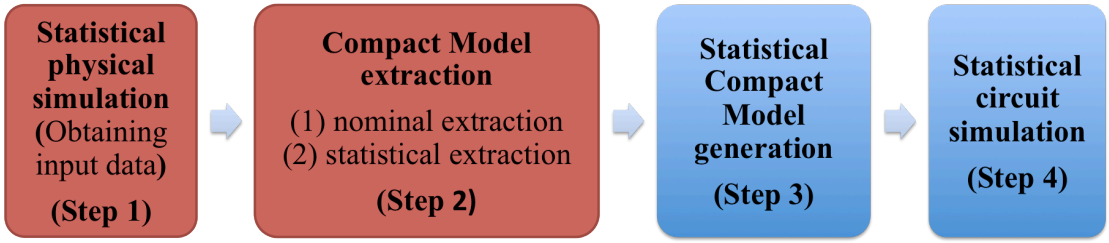


Fig.4.1 Step 1 and step 2 of the research flow.

Chapter 3 introduced the research flow and methodologies used in this thesis, with the exception of the generation of compact models which will be introduced in Chapter 5. This chapter focuses on the analysis of both the results from physical simulation (obtained upon completion of step 1) and subsequent compact model extraction (step 2) using the methods described in Chapter 3 (shown in Fig.4.1). A 25 nm gate length MOSFET is used for all simulations and the corresponding template for physical simulations is introduced in section 4.2. Physical simulation scenarios are introduced in section 4.2 and all simulations of statistical variability and BTI are performed using the statistical device simulator GARAND. The physical simulation results are analysed in section 4.3 at both a statistical level and individual device level in order to relate effects of statistical variability and BTI-induced ageing to physical device properties.

With the input data ready, compact model extraction is executed. The extracted models are analysed and compared with physical simulations to assess their accuracy in section 4.4. We show that extracted compact models can accurately represent not only the figures of merit extracted from the original device simulation data, but also the correlations between them. The extracted parameter distributions are analysed in section 4.4.3. The result of the work in this chapter is the creation of a lookup table (LUT) based statistical compact model library with 1,000 n- and p-MOS transistors at multiple discrete levels of degradation. The distribution of parameters extracted in stage 2 of the compact model extraction is monotonic. The statistical compact model extraction lays the ground-work for the statistical model generation strategy that will be outlined in Chapter 5.

4.2 Physical Simulation Scenario

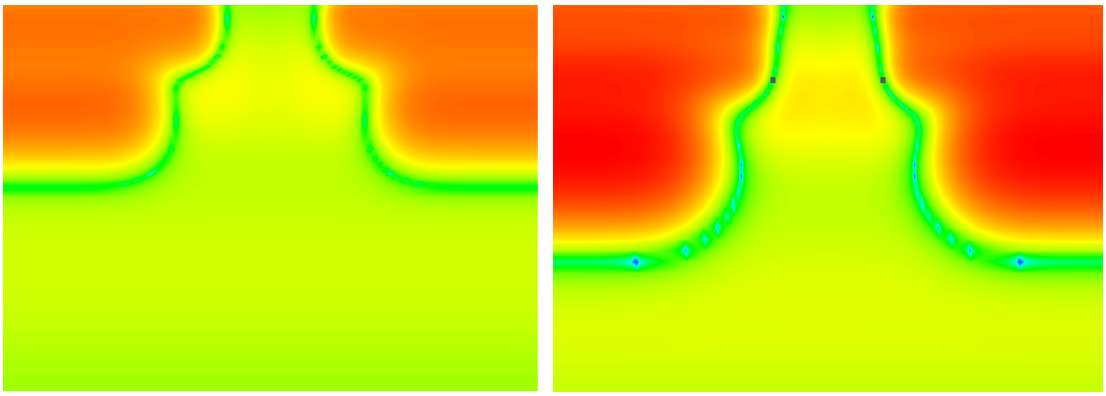
4.2.1 Compact Model Extraction Input Data

The effectiveness of large-scale statistical circuit simulations relies heavily on the accuracy of the statistical transistor models employed in SPICE simulation. It is common practice to make simplifying assumptions about the nature of the statistical variation, such as to assume a Gaussian distribution of threshold voltage, however in doing so accuracy is lost when these assumptions do not hold true. In order to develop truly accurate compact models for SPICE simulation it is necessary to first understand the distribution of device performances that are to be captured and to design a strategy that reflects this accurately in the model. In order to accurately understand the distribution of device performance, large-scale statistical device data is required. The statistical data should be seen separately from the nominal, or ‘designed’ transistor performance, as discussed in Chapter 3. Instead, the statistical data should be seen as an extension to the nominal performance, describing how sensitive the device is to internal physical variation. Following the extraction of a compact model to describe the nominal performance in Chapter 3, in this chapter, we focus on the extraction of statistical compact models designed to capture not only the effects of statistical variability, but of ageing and their interdependence as well.

The statistical device data used to develop the compact models is supplied by accurate statistical device simulation using the commercial statistical device simulator GARAND.

4.2.2 Device Description

In this study a 25 nm gate length bulk MOSFET representative of the 20 nm technology generation is used. The device follows the prescriptions of the ITRS-2010 update and is subject to realistic physical constraints. The device features a high-κ dielectric gate stack with 0.85 nm EOT and has a metal gate. The nominal p-channel device is designed to offer electrical characteristics that are as symmetrical as possible with that of the corresponding nominal n-channel devices.The device structures and doping profile are shown in Fig.4.2 and the important geometric and electrical parameters are summarised in Table 4.1. Full electrical transfer characteristics for the n- and p-channel devices are shown in Fig.4.3 and Fig.4.4, respectively.



(a) (b)

Fig.4.2 Net doping profiles for the template (a) n-channel 25 nm MOSFET and (b) p-channel 25 nm MOSFET. Used with permission.

Table 4.1 Structural and electrical parameters for the 25 nm n- and p- MOSFETs

Parameter		n-MOS	p-MOS	Description
L_g	[nm]	25	25	Physical gate length
EOT	[nm]	0.85	0.85	Equivalent Oxide Thickness
X_j	[nm]	15	22.5	Source/Drain Extension
N_A	[E18/cm^3]	4.5	4.95	Channel Doping
V_{dd}	[V]	1	1	Supply voltage
I_{off}	[nA/µm]	100	100	Off current
I_{on}	[µA/µm]	1351	1009	Drive current
Spacer	[nm]	24	24	

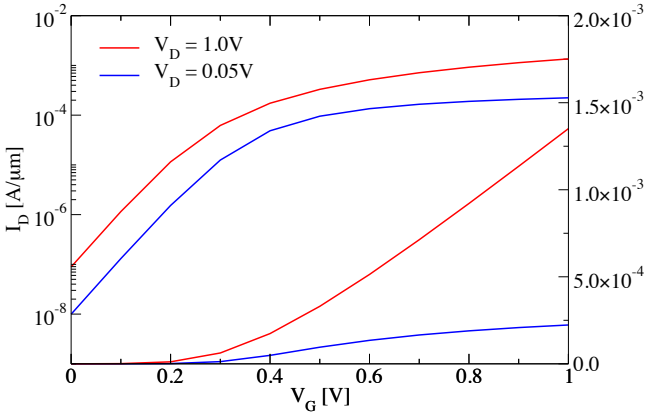


Fig.4.3 Transfer characteristics of the n-channel 25 nm template bulk MOSFET.

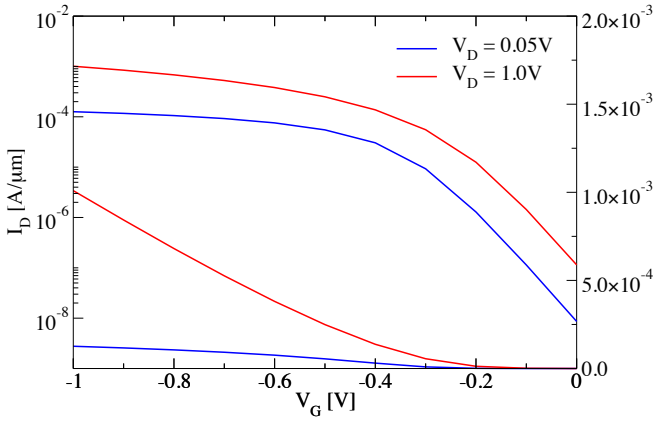


Fig.4.4 Transfer characteristics of the p-channel 25 nm template bulk MOSFET.

4.2.3 Simulation Scenario

With the device template introduced in section 4.2.2, physical simulations can be carried out. Simulations are carried out in order to resolve statistical variation in device performance associated with local variability as well as BTI-induced ageing. In this work, BTI-induced ageing is studied. The BTI related traps are present post manufacture, but can also be generated during circuit operation, and the number of occupied traps increases the longer the device remains in BTI conditions. BTI-induced ageing is represented accurately within these simulations by physically including occupied traps in the simulation domain. The traps also exhibit statistical variation as they vary from device to device in both trap numbers and positions. Importantly, the influence of a trap located at a certain position will be shown to be strongly dependent upon the local device variation, as this variation imposes additional sensitivity on device operation that is otherwise not resolved. In this work, simulations were carried out for the newly produced devices with no traps ("fresh" device), as well as for "young" devices at a low level of ageing (trap density: 1×10^{11} cm^{-2}), "middle-aged" devices at trap density of 5×10^{11} cm^{-2}, and "old" devices at trap density of 1×10^{12} cm^{-2}. Devices at all of the above stage of degradation, 'ages', are simulated in conjunction with statistical variability effects (RDD, LER, MGG). At each ageing level, simulations are performed at low drain bias (0.05V) and high drain bias (1V) separately. The simulation approach does not include transient analysis (for example, trapping or de-trapping events) but is designed to look at the statistics of a particular snapshot in time.

In order to decide the ensemble size of devices for each trap density, standard errors of mean, standard deviation, skew and kurtosis with different sample sizes are estimated. These moments are used to describe the statistical properties of a sample and understanding the error associated with each allows an estimation of the ensemble size needed to obtain a certain accuracy in each of these moments. The following equations show the standard errors when using different sample size to estimate the mean, standard deviation, skewness and kurtosis of the population, in which n accounts for the sample size [95].

$$\sigma_{mean} = \frac{\sigma_{sample}}{\sqrt{n}} \tag{4.1}$$

$$\sigma_{standard\ deviation} \approx \frac{1}{\sqrt{2(n-1)}} \tag{4.2}$$

$$\sigma_{skewness} \approx \sqrt{\frac{6}{n}} \tag{4.3}$$

$$\sigma_{kurtosis} \approx \sqrt{\frac{24}{n}} \tag{4.4}$$

Table 4.2 shows the calculated standard errors using equations (4.1)-(4.4). Considering the computational time and accuracy, the ensemble size of 1,000 devices at each trap density is chosen in the generated statistical data. Such simulations take approximately two working weeks on the state-of-art computational cluster and led to significantly higher accuracy than the available commercial tools. The physical simulation scenarios are shown in Fig.4.5.

Table 4.2 Standard errors of mean, standard deviation, skewness and kurtosis with different sample size.

Sample size (n)	Standard errors of mean	Standard error of standard deviation	Standard error of skewness	Standard error of kurtosis
100	10%	7.11%	24.49%	48.99%
200	7.07%	5.01%	17.32%	34.64%
500	4.47%	3.17%	10.95%	21.91%
1000	3.16%	2.24%	7.75%	15.49%
2000	2.24%	1.58%	5.48%	10.95%
5000	1.41%	1.00%	3.46%	6.93%
10000	1.00%	0.71%	2.45%	4.90%

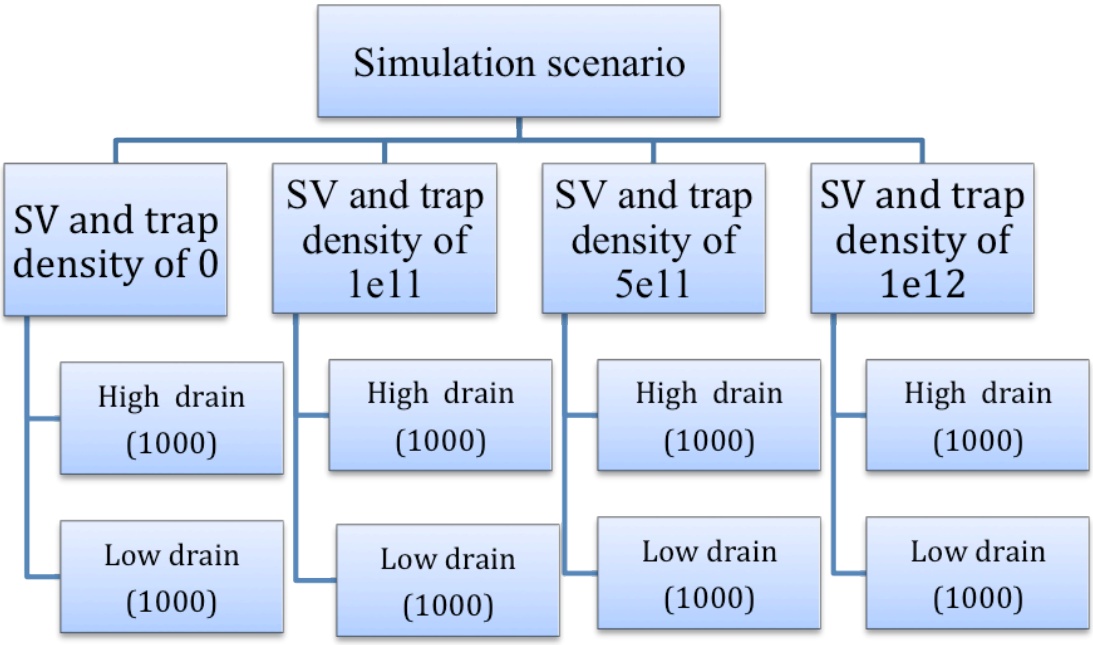


Fig.4.5 Physical simulation scenario.

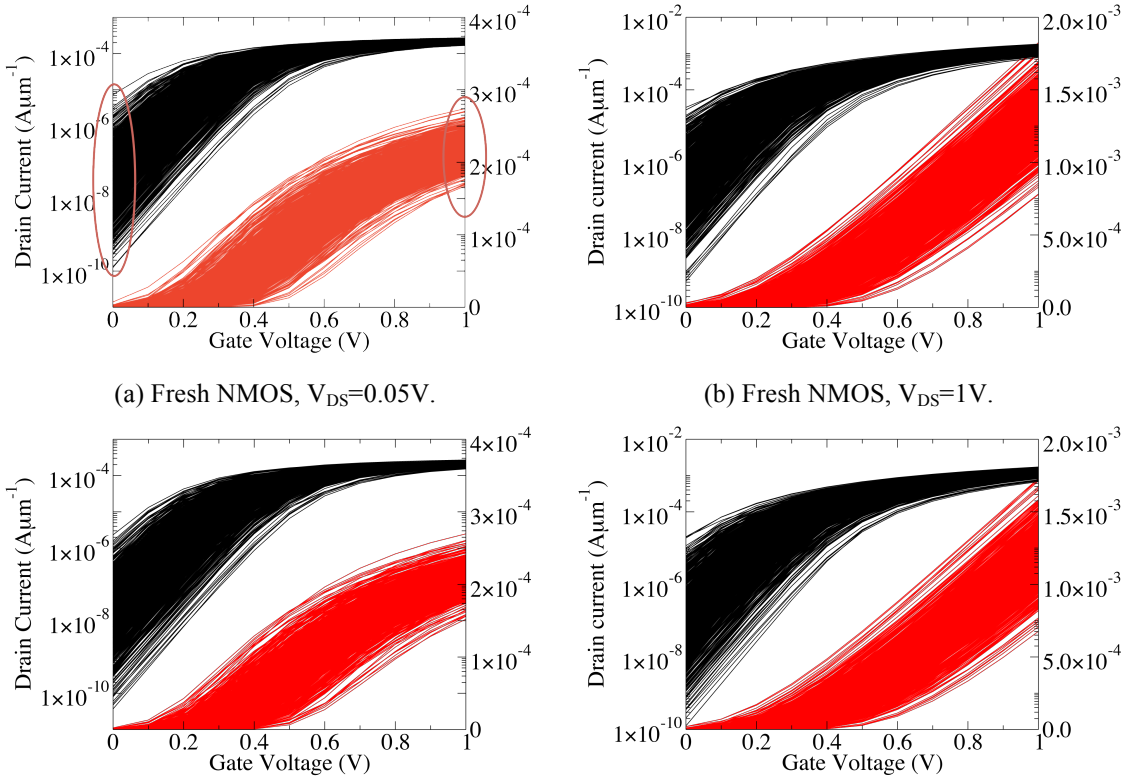
4.3 Physical Simulation Results and Discussion

Physical simulations are performed by scenarios stated above using the GSS 'atomistic' TCAD software – 'GARAND'. The simulation results are investigated in this section for a deep understanding of device behaviors under statistical variability and at different levels of ageing. The results will be used as input data for the second stage of compact model extraction, and as reference data for the comparison through the whole research.

4.3.1 I_DV_G Curves Comparisons.

Full I_DV_G curves for each device of the ensembles discussed in the last section are simulated and devices at trap density of 0 and $1x10^{12}cm^{-2}$ are shown in Fig.4.6 and Fig.4.7, at low drain (0.05V) and high drain bias (1V) respectively. It can be seen in these pictures that statistical variability induced variations in the current voltage characteristics are non-negligible, that the magnitude of off-current variations spreads over 5 orders of magnitude, while the on-currents are widely distributed as well. These variations become even worse as trap density increases. The big variations of device performances can lead to uncontrollable circuit behaviours, for example, poorly matched threshold voltages which diverge over age can result in disasters in high precision analog

circuits, in which the extremely stable threshold voltage is required over the circuit life time [96]. Though device performances varies greatly due to statistical variability, ageing exaggerates the variation of device behaviours to different degrees according to different ageing levels, which can be seen in the distributions of figures of merit in 4.3.2. I_DV_G curves give a direct and broad view of the devices' performances. The widely spread distributions of the I_DV_G curves emphasize the significance of the effects of variability and ageing. This emphasises the importance of investigating these intrinsic variations at different levels of ageing, and the trend of device performances when ageing increases.



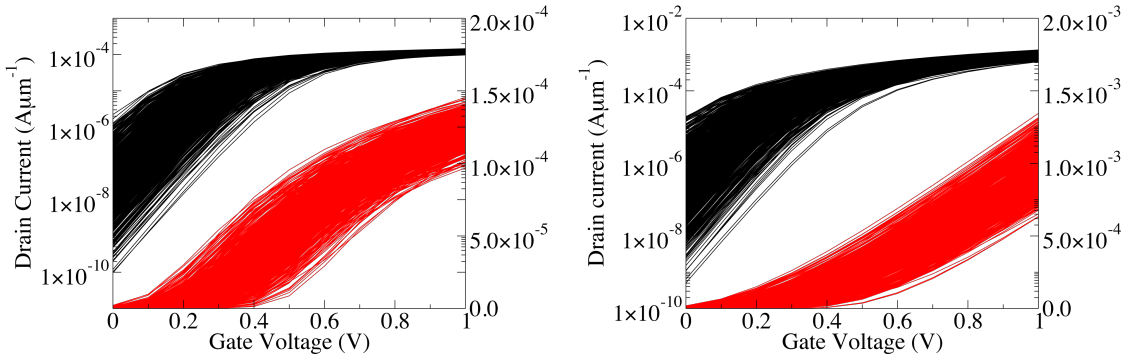
(a) Fresh NMOS, V_{DS}=0.05V.

(b) Fresh NMOS, V_{DS}=1V.

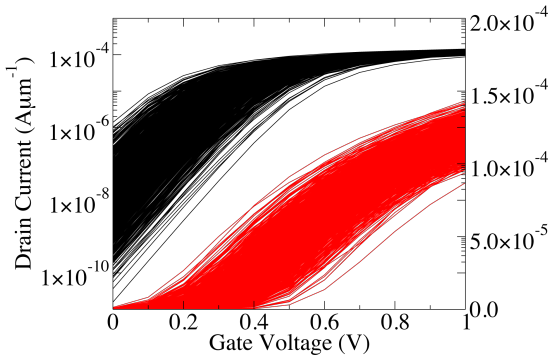
(c) NMOS at trap density of $1\times10^{12}cm^{-2}$, V_{DS}=0.05V.

(d) NMOS at trap density of $1\times10^{12}cm^{-2}$, V_{DS}=1V.

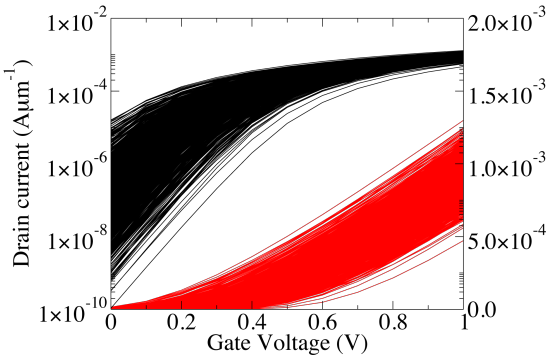
Fig.4.6 Current voltage characteristics of NMOS devices.



(a) Fresh PMOS, V_{DS}=0.05V.



(b) Fresh PMOS, V_{DS}=1V.



(c) PMOS at trap density of $1\times10^{12}cm^{-2}$, V_{DS}=0.05V.

(d) PMOS at trap density of $1\times10^{12}cm^{-2}$, V_{DS}=1V.

Fig.4.7 Current voltage characteristics of PMOS devices.

4.3.2 Figures of Merit Comparisons.

The key device figures of merit (V_{TH}, I_{ON}, I_{OFF}, DIBL) in respect to MOSFETs and their significances were discussed in Chapter 3. We can better understand transistor performance by studying the distributions of the key figures of merit as they play the role of the performance index of each device. V_{TH} defines the voltage at which the device turns on, while I_{ON} and I_{OFF} are the drive current and leakage current respectively. DIBL defines the difference in V_{TH} at low and high drain bias and is a good first-order indication of short channel effects. Increased trap charges screen the gate potential from the channel and reduce the inversion charge density and consequently increase the threshold voltage V_{TH}. However, associated with increased trap density in statistical device simulation, it is a higher probability that a trapped charge increases the potential barrier in an area that has a low barrier due to the distribution of discrete dopants. A single trap may therefore effectively cut-off the current flowing through a percolation path between source and drain and give rise to a very large increase in V_{TH} far larger than that associated with the reduction in overall inversion charge.

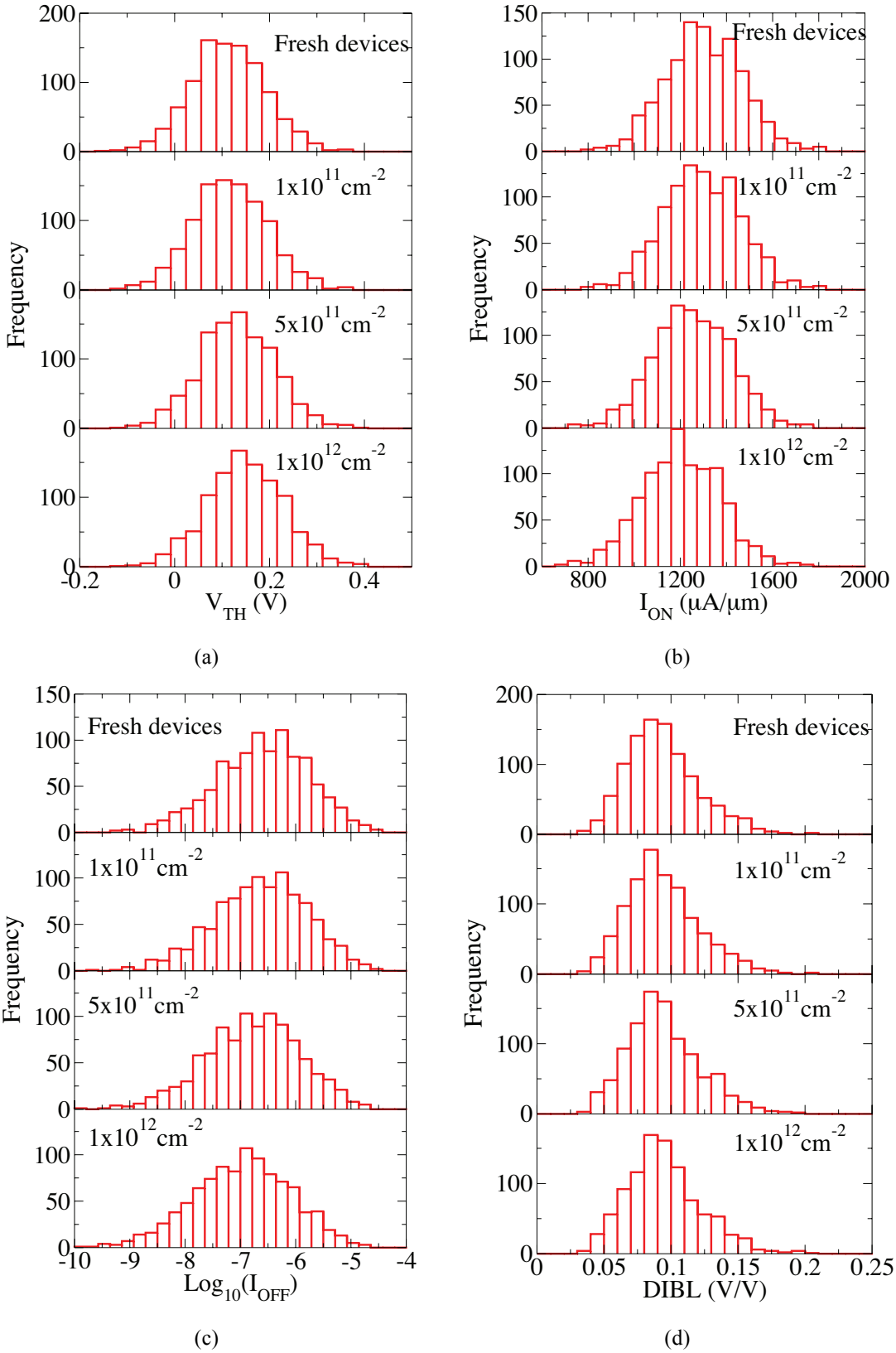


Fig.4.8 NMOS V_{TH} (a), I_{ON} (b), I_{OFF} (c) and DIBL (d) distributions, when Vd=1V. From the top to the bottom pictures, trap densities are 0, $1x10^{11}cm^{-2}$, $5x10^{11}cm^{-2}$, and $1x10^{12}cm^{-2}$ respectively.

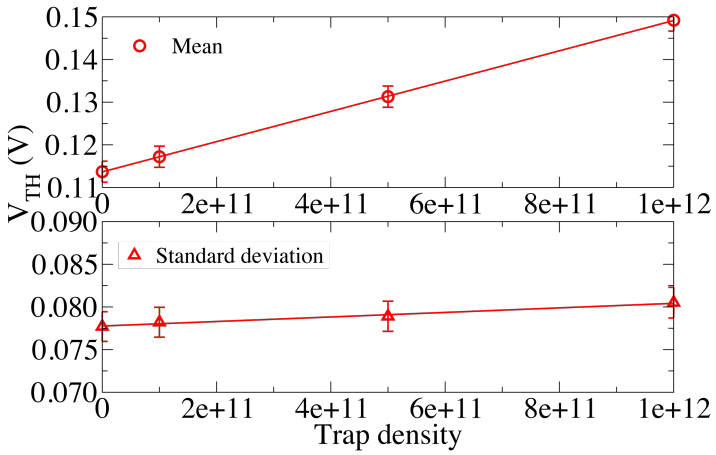


Fig.4.9 Mean and standard deviation of V_{TH} for NMOS, at V_{DS}=1V.

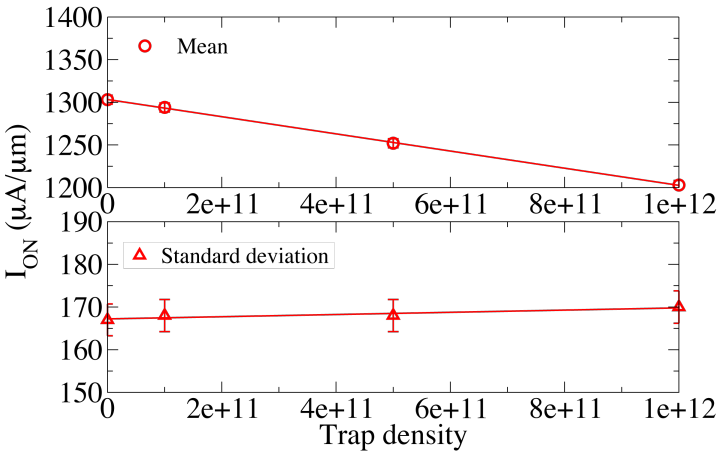


Fig.4.10 Mean and standard deviation of I_{ON} for NMOS, at V_{DS}=1V.

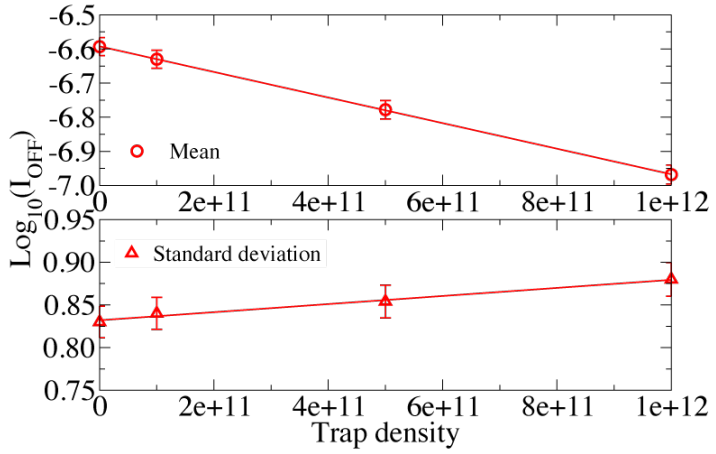


Fig.4.11 Mean and standard deviation of Log10(I_{OFF}), at V_{DS}=1V.

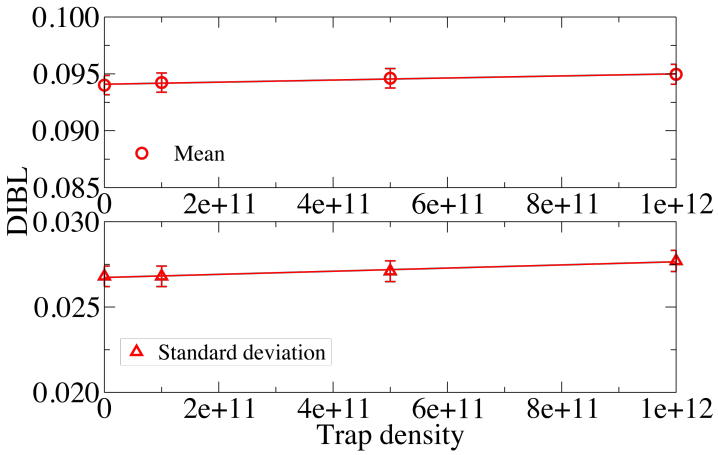


Fig.4.12 Mean and standard deviation of DIBL, at V_{DS}=1V.

Fig.4.8 shows the distribution of NMOS devices figures of merit at high drain bias. Fig.4.9 to Fig.4.12 show the corresponding mean and standard deviation of the distributions. It is clear that V_{TH} increases with the increase of trap density. Compared with fresh devices and in line with the increase in V_{TH}, I_{ON} and I_{OFF} for NMOS devices at high drain bias decrease as trap density increases. These results match the expectation stated above. Since statistical variability and ageing effect influences V_{TH} at both high drain bias and low drain bias, DIBL does not change significantly. This agrees with the physical simulation DIBL results of NMOS devices shown in Fig.4.8 (d). The standard deviation of each figure of merit increases slightly as trap density increases, indicating that ageing brings more variations and reflecting the wider distributions of the I_DV_G curves discussed in section 4.3.1.

QQ plots for NMOS at low drain bias, PMOS at high drain bias and low drain bias are shown from Fig.4.13 to Fig.4.15. The descriptive statistics of NMOS and PMOS are shown in appendix A. These data show the same trend as ageing level increases.

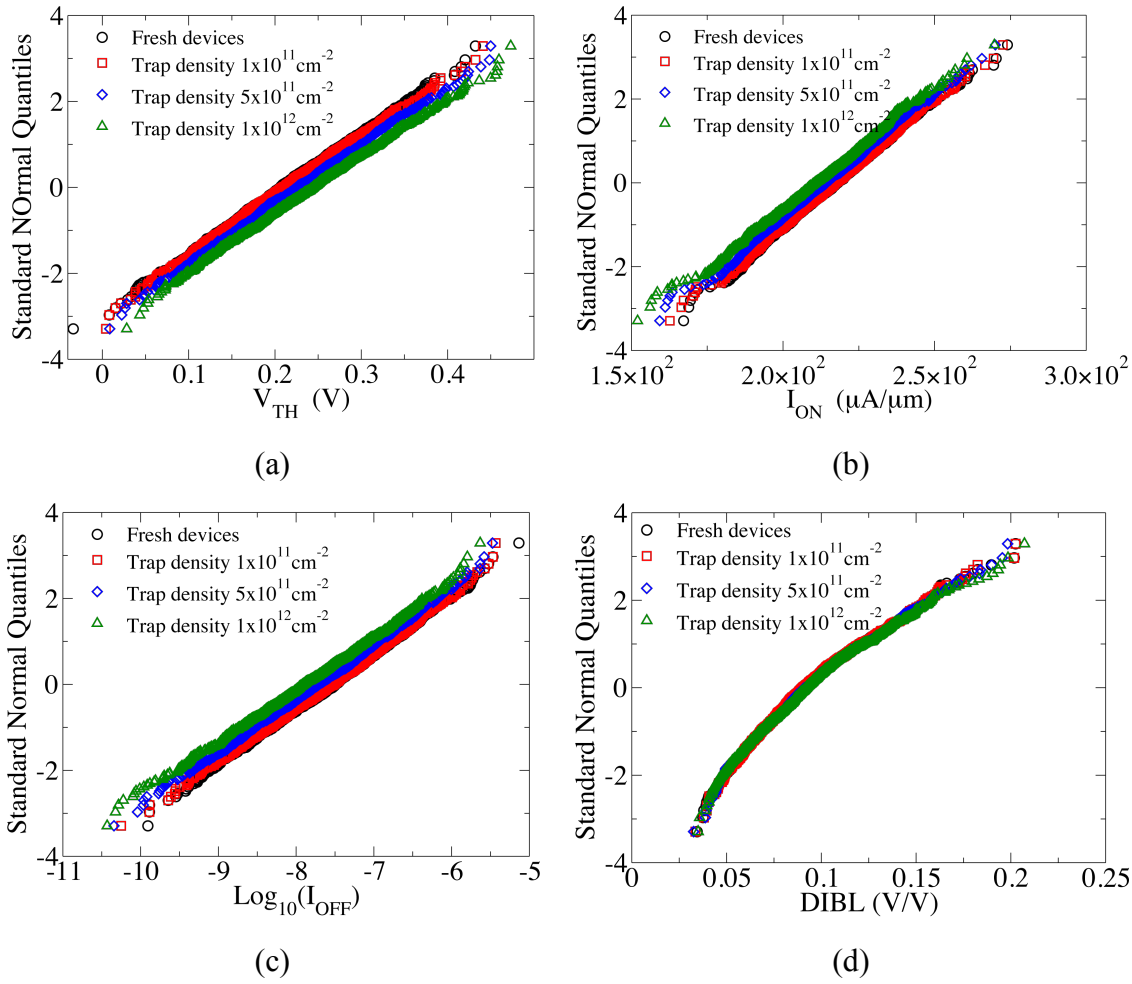
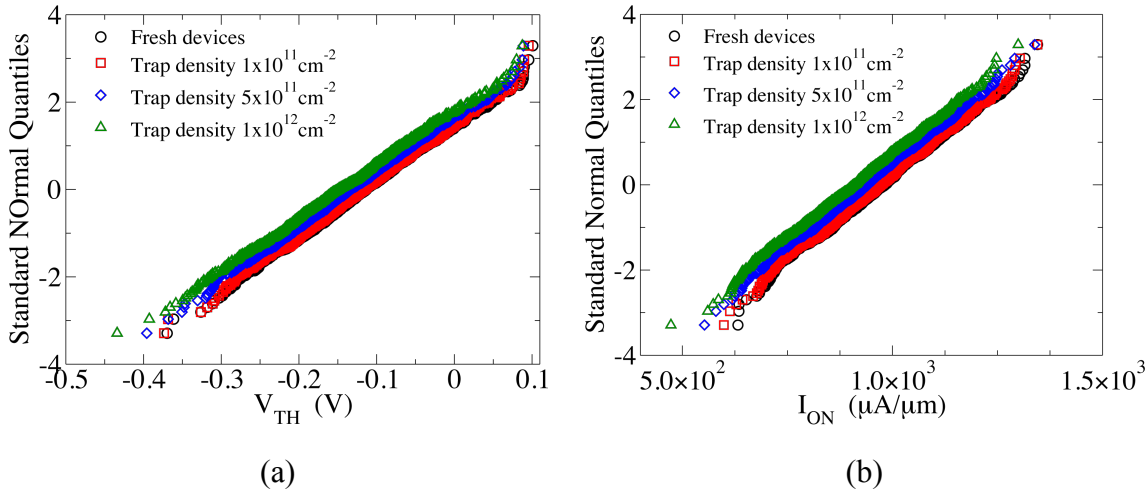
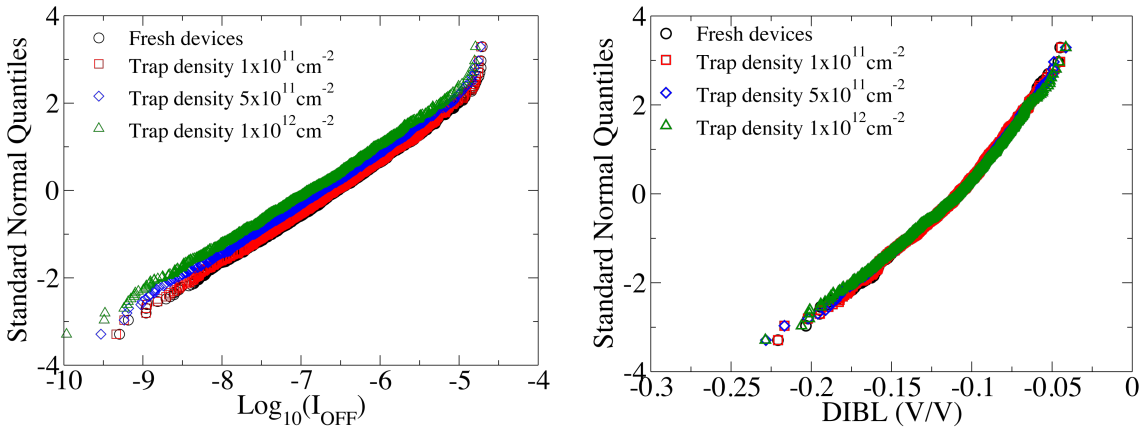


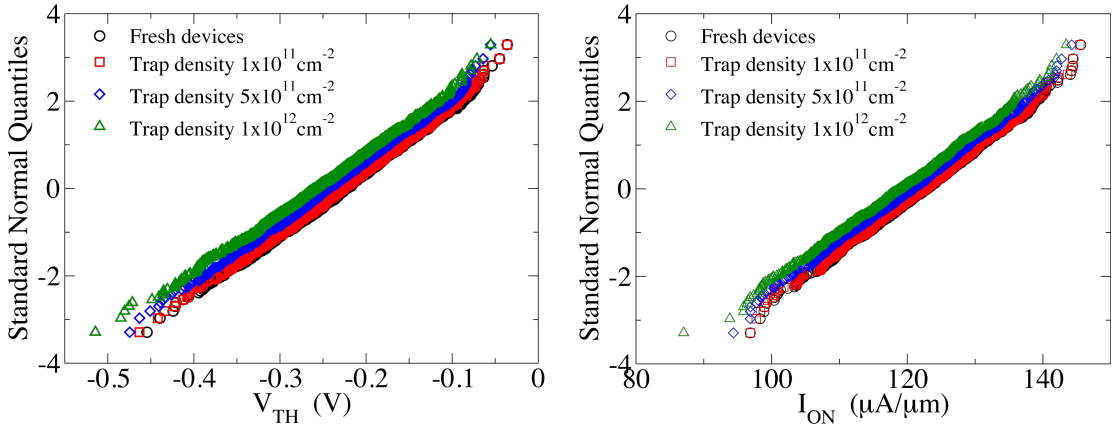
Fig.4.13 QQ plots of figures of merit for NMOS devices, V_{DS}=0.05V.



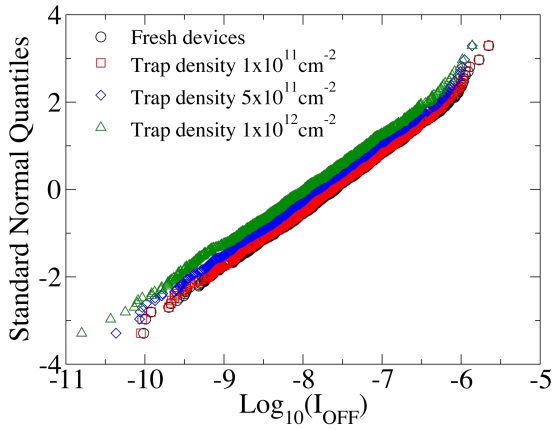


(c) (d)

Fig.4.14 QQ plots of figures of merit for PMOS devices, V_{DS}=1V.



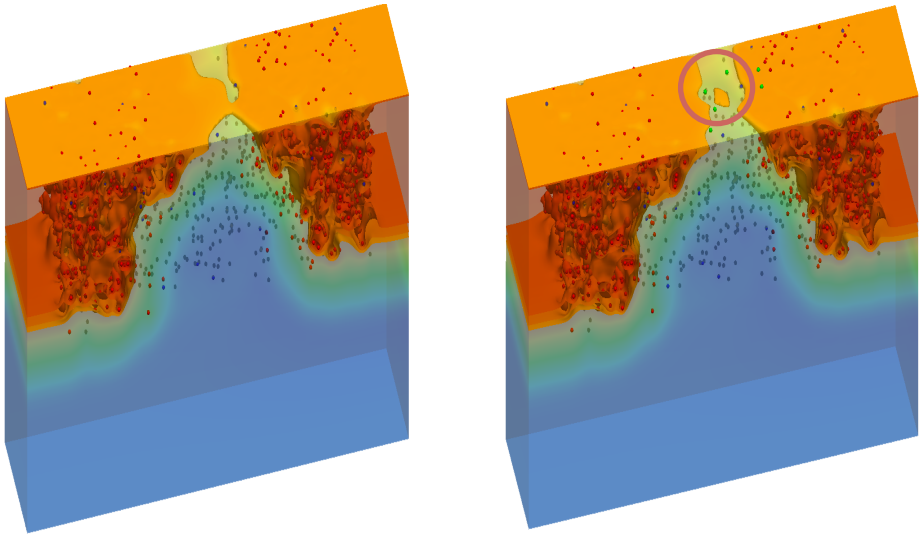
(a) (b)



(c)

Fig.4.15 QQ plots of figures of merit for PMOS devices, V_{DS}=0.05V.

4.3.3 Individual Device Performance.



(a) (b)

Fig.4.16 The electron density in a single atomistically simulated device (Device No.136).

The large statistical sample of simulated devices allows the influence of statistical variability and BTI-induced ageing on device performance to be analysed. For an individual device, trapping of a single discrete charge in the vicinity of a current "percolation path" can result in a dramatic change in the transistor characteristics, including a large threshold voltage shift and on-current reduction. Among the 1,000 simulated devices, No. 136, which is shown in Fig.4.16, is taken as an example because this device shows the maximum V_{TH} increase of 145mV. Fig.4.16 (a) shows this device at fresh situation and a clear current path exists between the source and drain. Fig.4.16 (b) shows this device at trap density of $1x10^{12}cm^{-2}$, and the trapped charge (shown in green) has completely closed off the current path resulting in a large change in device threshold voltage. If the trapped charge had appeared at different place, the corresponding V_{TH} shift may have been much smaller. Investigating into individual device performance and the corresponding structure shows the stochastic nature of the influence by statistical variability and BTI-induced ageing. This also illustrates the necessity of large statistics for the investigation of statistical variability and ageing, no matter for the device performance or for the circuit performance.

4.3.4 The Analysis of ΔV_{TH} and ΔI_{ON}.

From the previous section, we can see that with trap density increases, device V_{TH} increases and on-current decreases. Fig.4.17 shows ΔV_{TH} (V_{TH} of device at particular trap density minus V_{TH} of fresh device) and Fig.4.18 shows ΔI_{ON} (I_{ON} of fresh device minus I_{ON} of device at particular trap density) respectively. These figures clearly present the distributions of the increase of V_{TH} and decrease of I_{ON} for each simulated ageing level. Fig.4.19 shows the scattering plot of V_{TH} and I_{ON}, indicating the de-correlation between V_{TH} and I_{ON}. The underlying reason is that when traps block the current "percolation path", threshold increases. However, as bias increases on gate stack, other current "percolation path" may form depending on the dopants and traps distribution, showing the stochastic nature of statistical variability and ageing. Therefore, I_{ON} shift is not highly correlated with V_{TH} shift.

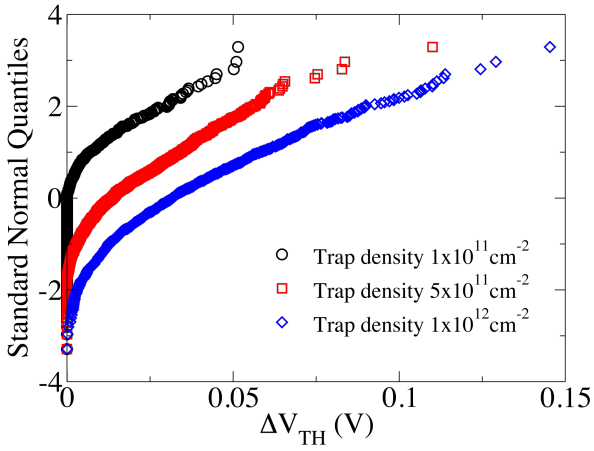


Fig.4.17 ΔV_{TH} distribution for NMOS when V_{DS}=1V.

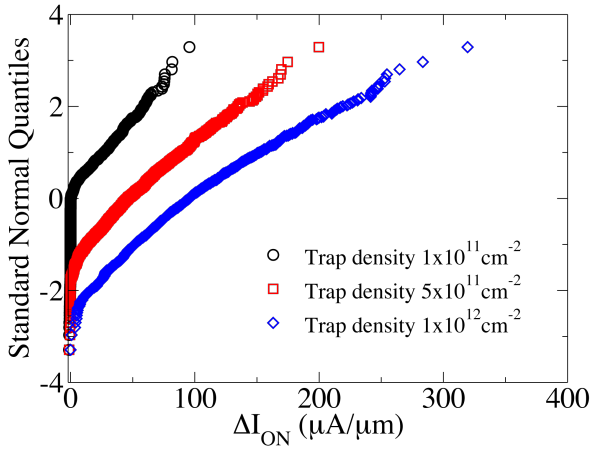


Fig.4.18 ΔI_{ON} distribution for NMOS when V_{DS}=1V.

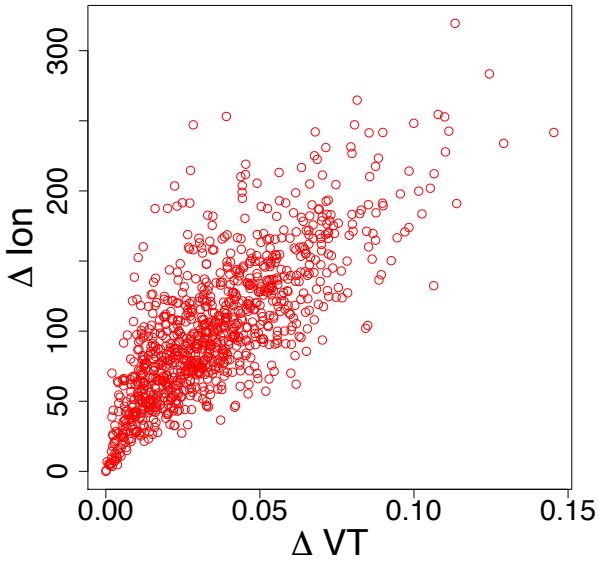


Fig.4.19 The scattering plot showing the correlation between ΔV_{TH} and ΔI_{ON} at trap density of $1x10^{12}cm^{-2}$ when V_{DS}=1V.

Physical simulations give clearly transistor performances under the influence of statistical variability and by levels of ageing (0, low, medium and high). As device ageing increases, the threshold voltage increases. Device off-current and on-current decreases. As ageing becomes more and more severe, it will be harder to turn on the device and device performance will be lower due to the reduced drive current. In the end, the dielectric layer may brake down and the device fail completely. The physical simulation data will be used as the resource data in the second stage for the compact model extraction.

4.4 Compact Model Extraction Results

4.4.1 Statistical Compact Model Extraction

The two-stage compact model extraction methodology used in this section is outlined in detail in Chapter 3, along with the seven carefully selected parameters that are used to describe simulated device behaviour under the influence of statistical variability and ageing. These seven parameters are used in the second stage, figure of merit based statistical compact model extraction. For each simulated device from the ensembles at each device 'age', the seven parameters are re-extracted using the methodology in

section 3.3.2. Then each set of seven parameters for each particular device is inserted into the uniform compact model to constitute the look-up table models. In the next section, extracted compact models are compared against the physical simulation to verify the accuracy of this extraction methodology. The seven re-extracted parameters' distributions are analysed in section 4.4.3 as well.

4.4.2 Comparisons Between Physical Simulation Results and Extracted Compact Models

In order to verify the accuracy of the extracted compact models in reproducing the device performance, they are compared against the physical simulation results upon which they are based. The extraction errors for NMOS and PMOS are shown in Fig.4.20 and Fig.4.21. From these figures we can see that nearly all the errors are below 6%, except for a small tail beyond 6%. Considering the fact that errors in the subthreshold can artificially inflate the percentage error because they're exponential, it is reasonable to conclude that the error figures show very positive results of the compact model extraction accuracy.

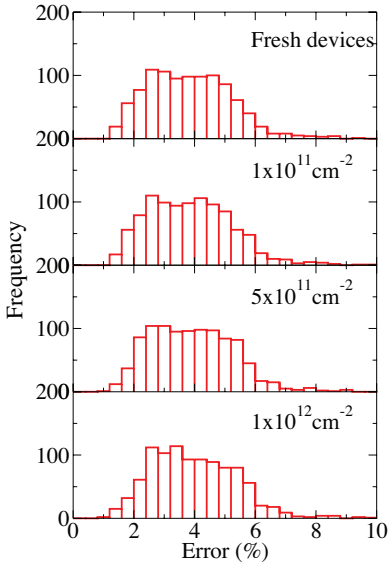


Fig.4.20 The error distribution of the compact model extraction for NMOS.

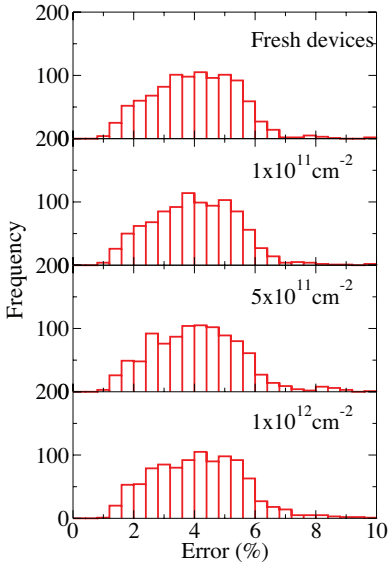


Fig.4.21 The error distribution of the compact model extraction for PMOS.

Aside from verifying the low value for the extraction errors, it is important that the extracted compact models accurately capture the distributions of the critical device figures of merit and the correlations between these figures of merit from physical simulations. Fig.4.22 and Fig.4.24 show the distributions of V_{TH}, I_{ON}, I_{OFF} and DIBL of fresh devices and Fig.4.23 and Fig.4.25 show that of devices at trap density of $1x10^{12}cm^{-2}$. The black circles are the results from physical simulations while the red squares represent figures of merit from extracted compact models. Comparing the results, we can see that no matter if the devices are with statistical variability only (fresh devices) or highly degraded (devices at trap density of $1x10^{12}cm^{-2}$), the extracted compact models can capture the distributions of the figures of merit from physical simulations. This verifies the accuracy and physical validity of the figure of merit based extraction methodology.

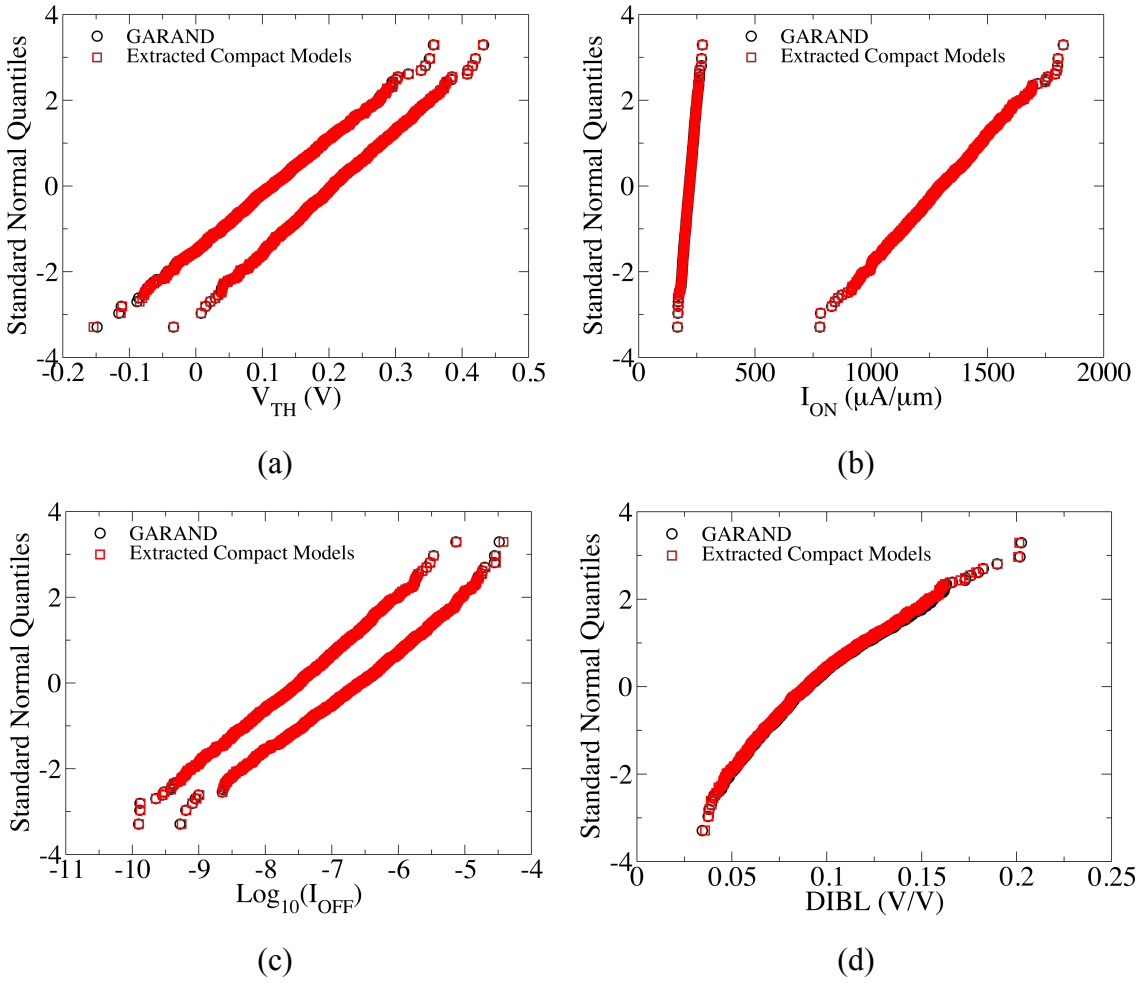


Fig.4.22 The comparisons of V_{TH}, I_{ON}, $Log_{10}(I_{OFF})$, DIBL for fresh NMOS devices, between physical simulation results and extracted compact models at high drain and low drain bias.

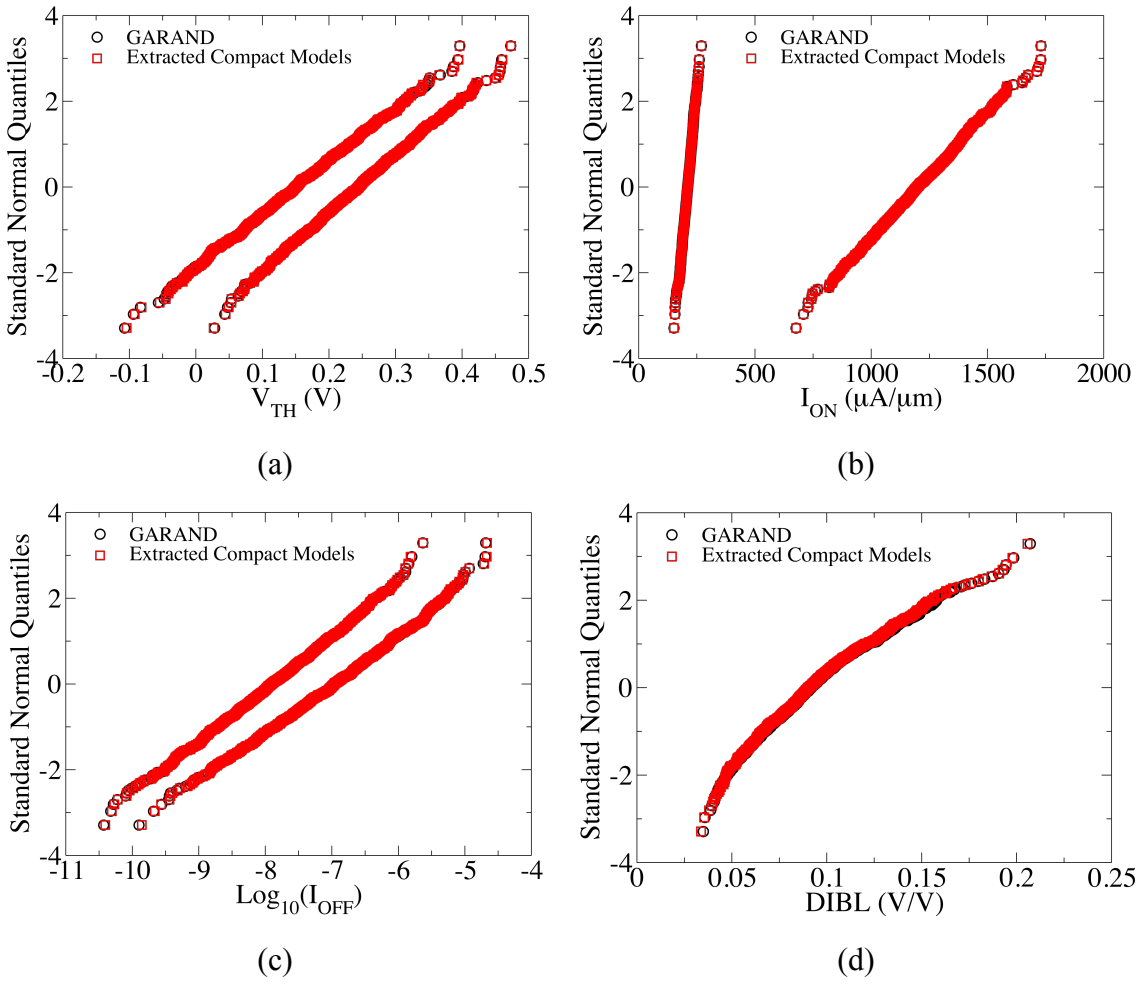
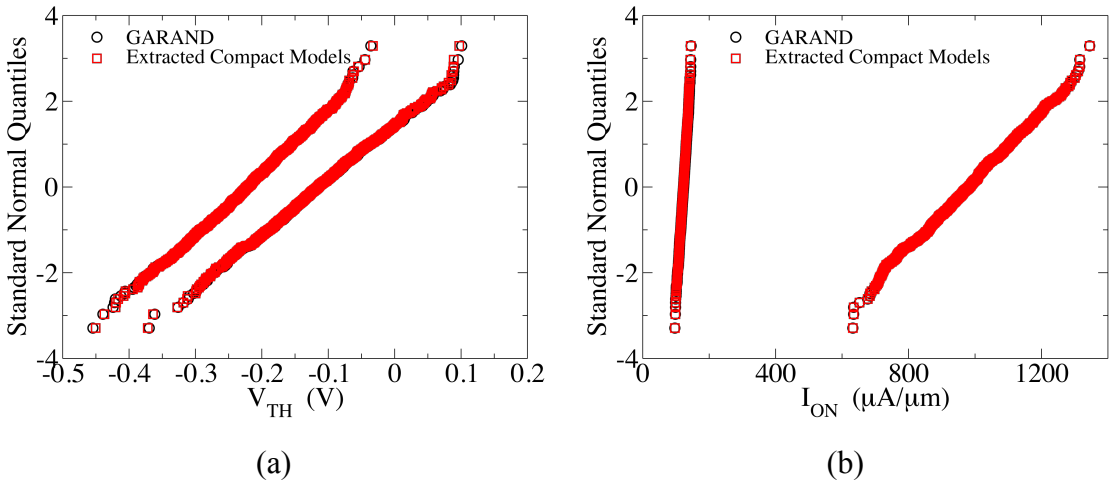
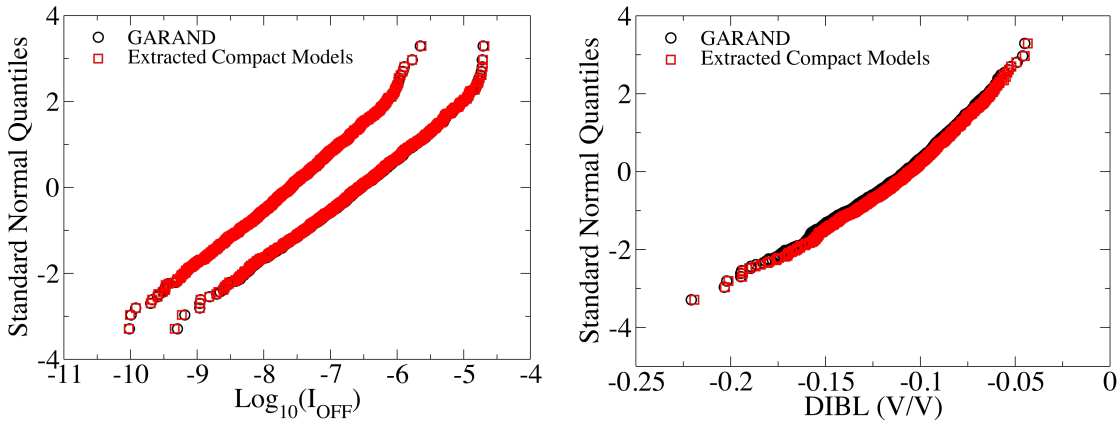


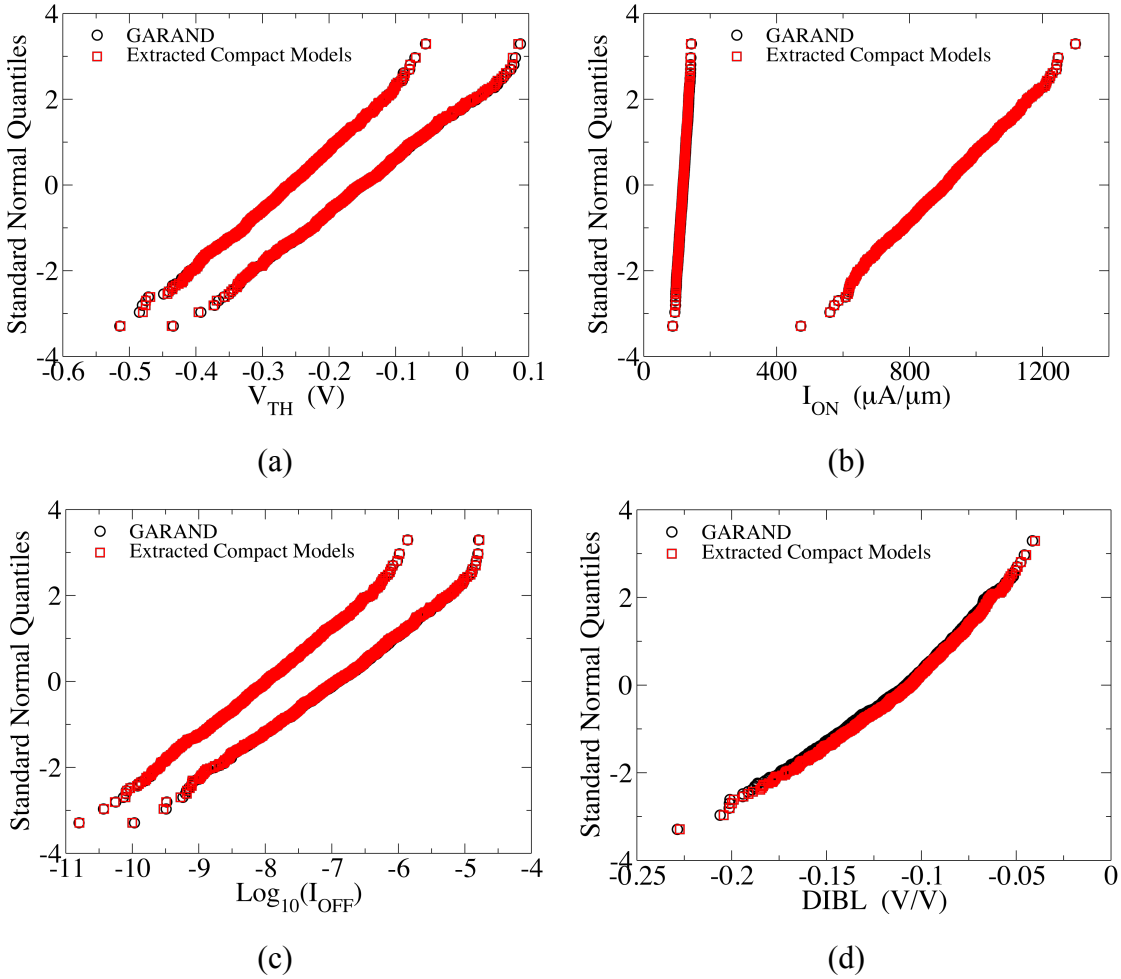
Fig.4.23 The comparisons of V_{TH}, I_{ON}, $Log_{10}(I_{OFF})$, DIBL for NMOS devices at trap density of $1x10^{12}cm^{-2}$, between physical simulation results and extracted compact models at high drain and low drain bias.





(c) (d)

Fig.4.24 The comparisons of V_{TH}, I_{ON}, $Log_{10}(I_{OFF})$, DIBL for fresh PMOS devices, between physical simulation results and extracted compact models at high drain and low drain bias.



(a) (b)

(c) (d)

Fig.4.25 The comparisons of V_{TH}, I_{ON}, $Log_{10}(I_{OFF})$, DIBL for PMOS devices at trap density of $1x10^{12}cm^{-2}$, between physical simulation results and extracted compact models at high drain and low drain bias.

The device figures of merit are correlated with each other because of the underlining physics of the device. Aside from capturing figures of merit, the extracted compact models capture the correlations between figures of merit from physical simulations, indicating that the compact models successfully capture the physical behaviour of the device under local variation. Fig.4.26 to Fig.4.29 compare the correlations of the figures of merit from physical simulations of fresh devices and devices at the highest ageing level (at trap density of $1x10^{12}cm^{-2}$), with those from the extracted statistical SPICE compact models. The black figures show the correlations of figures of merit from physical simulations while the red figures show the correlations of figures of merit from extracted compact models. All of the figures show excellent agreement with the physical simulations, capturing both the Pearson correlation coefficient and the form (shape) of the relationship, giving confidence that the compact models provide a good, physical, representation of the operation of each device.

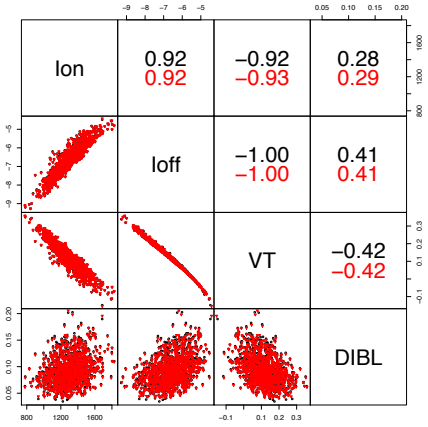


Fig.4.26 Correlations between figures of merit for fresh NMOS devices (V_{DS}=1V). The black indicates physical simulation results. The red indicates extracted compact model results.

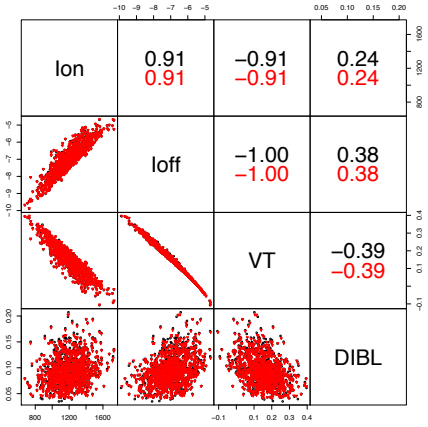


Fig.4.27 Correlations between figures of merit for NMOS devices at trap density of $1x10^{12}cm^{-2}$ (V_{DS}=1V). The black indicates physical simulation results. The red indicates extracted compact model results.

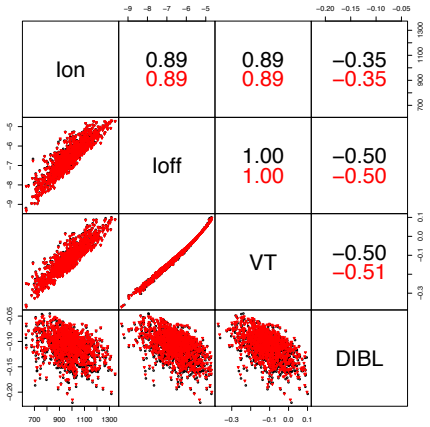


Fig.4.28 Correlations between figures of merit for fresh PMOS devices (V_{DS}=1V). The black indicates physical simulation results. The red indicates extracted compact model results

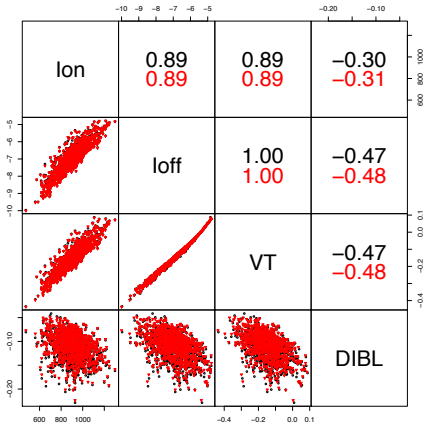


Fig.4.29 Correlations between figures of merit for PMOS devices at trap density of $1x10^{12}cm^{-2}$ (V_{DS}=1V). The black indicates physical simulation results. The red indicates extracted compact model results

4.4.3 Re-extracted Parameters

The seven statistically extracted parameters, VTH0, EAT0, VOFF, NFACTOR, UA, VSAT and CDSCD, are selected for the description of statistical device behaviour due to the importance they play in resolving the critical features associated with statistical variability and ageing. VTH0 captures V_{TH} at low drain bias and ETA0 captures DIBL. VOFF captures the effect of off-current and UA accounts for the mobility variation due to variation in dopant numbers and position. NFACTOR and CDSCD capture the low drain and high drain subthreshold slope separately. VSAT is the velocity saturation parameter that used to capture the variations of I_{ON}. Although these parameters were selected for this study, their ability to capture statistical variability and reliability effects is heavily dependent upon the underlying technology and the quality of the uniform compact model.

It should be noted that though each of these parameters play an important role in capturing their own effect, they may contribute to capture each other's effect as well. This is due to the complex correlations between parameters in BSIM4. Fig.4.30 to Fig.4.33 show the scatter plots and correlations between the re-extracted parameters of

NMOS and PMOS at trap density of 0 and $1x10^{12}cm^{-2}$ respectively. From these figures we can see that the re-extracted parameters are obviously not statistically independent, some of which are highly correlated while some are not. These complex correlations show compact model's complex nature.

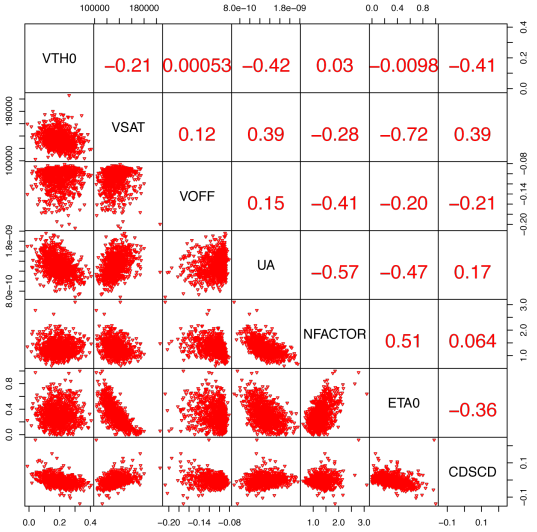


Fig.4.30 The extracted parameters' correlations of NMOS fresh devices.

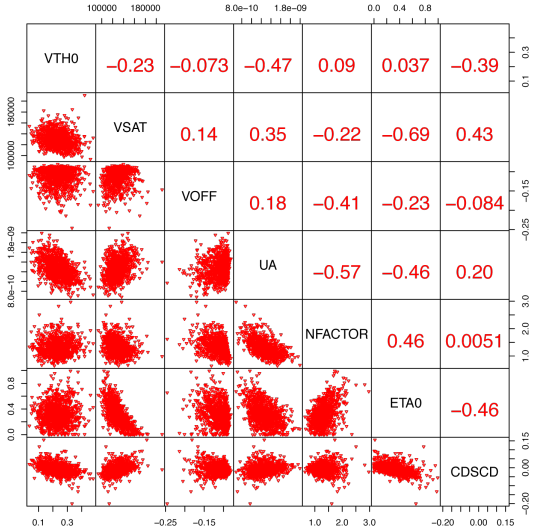


Fig.4.31 The extracted parameters' correlations of NMOS devices at trap density of $1x10^{12}cm^{-2}$.

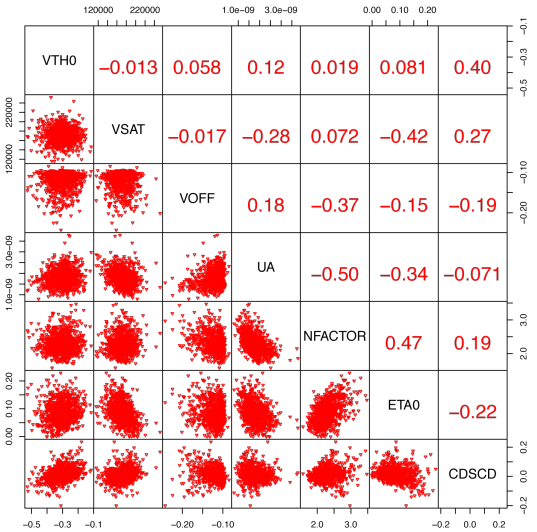


Fig.4.32 The extracted parameters' correlations of PMOS fresh devices.

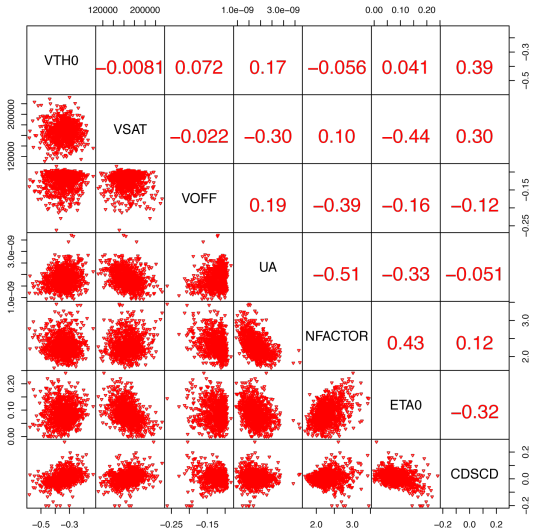


Fig.4.33 The extracted parameters' correlations of PMOS devices at trap density of $1x10^{12}cm^{-2}$.

The statistically extracted parameters' distributions are shown from Fig.4.34 to Fig.4.40. From these figures, it is clear that VTH0 shifts more than other parameters as the trap

density increases. As the first order effect of statistical variability and ageing, V_{TH} shift should be greatly reflected by the change of VTH0. As the ageing level increases, the mean value of VTH0 increases from 0.194 to 0.233 (20.1% increase) while the NMOS V_{TH} at low drain bias changes from 0.207 to 0.244 (17.87% increase), which are shown in Fig.4.13 (a) and Fig.4.34 respectively. Different amount of increase between VTH0 and V_{TH} also show that VTH0 contribute most to V_{TH} at low drain bias, but other parameters also contribute.

The distribution of VSAT, VOFF, and UA are shown from Fig.4.35 to Fig.4.37. With the increase of ageing levels, on current and off current are both decreased, which are directly reflected by VSAT and VOFF. The UA distributions mirror the decrease of mobility in Fig.4.37. This is a crucial result, as it shows that the mean VTH0 shift is not sufficient to capture the impact of BTI based degradation.

ETA0, which captures DIBL effect, nearly does not shift (shown in Fig.4.39) when ageing level increases. This is in agreement with the physical simulation and extracted compact model results (shown in Fig.4.13 (d)). NFACTOR and CDSCD, which are shown in Fig.4.38 and Fig.4.40, have wide distributions but do not show significant changes when trap density increases. This indicates that subthreshold slope is not changed by ageing but is greatly influenced by statistical variability.

The seven re-extracted parameters' moment distributions have the same characteristic that they are monotonic and linearly changed as the trap density increases, from which the compact model generation benefits (This will be demonstrated in Chapter 5). Historically, NBTI effects in PMOS dominated the ageing performance of SRAM cells due to the higher occupation of traps than in NMOS. However, following the introduction of high k metal gate technology, PBTI in NMOS was brought into focus and comparable to trapping in PMOS. In this study, application of trapping to NMOS will be discussed directly. Application to PMOS follows identically and results are placed in the appendix for completeness. The distributions of these parameters for PMOS devices show a similar result, which is shown in appendix B.

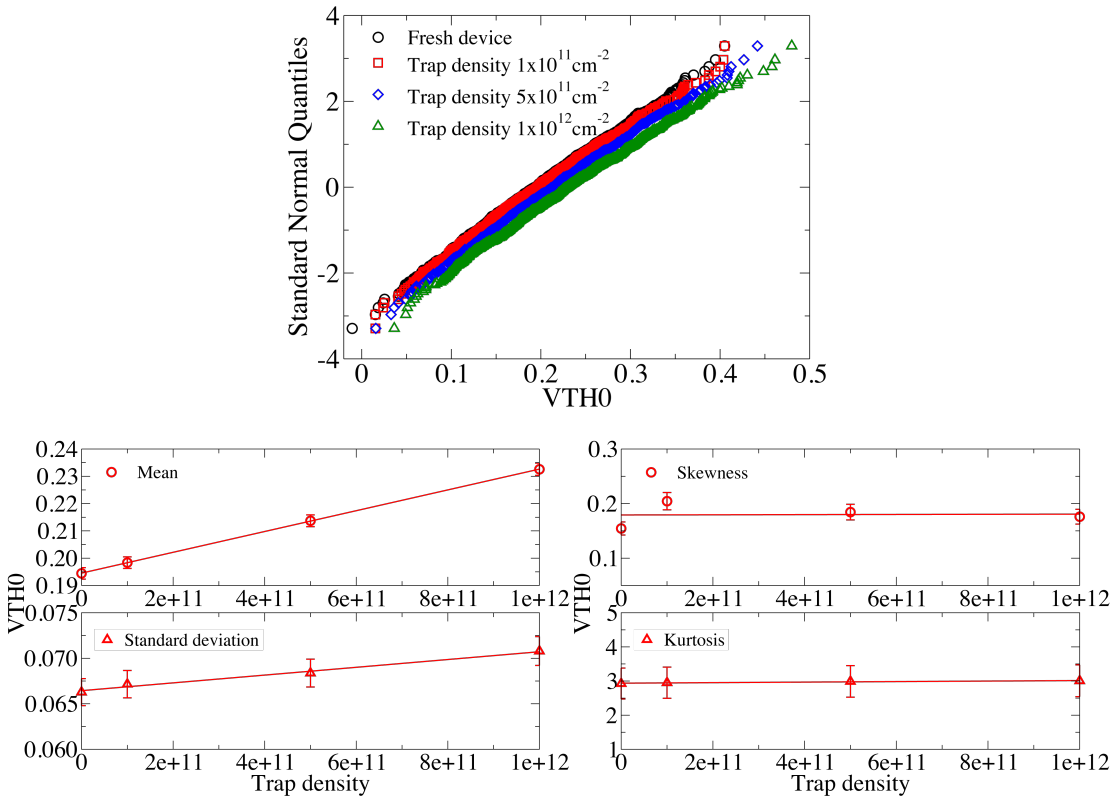


Fig.4.34 The distribution of VTH0 for NMOS.

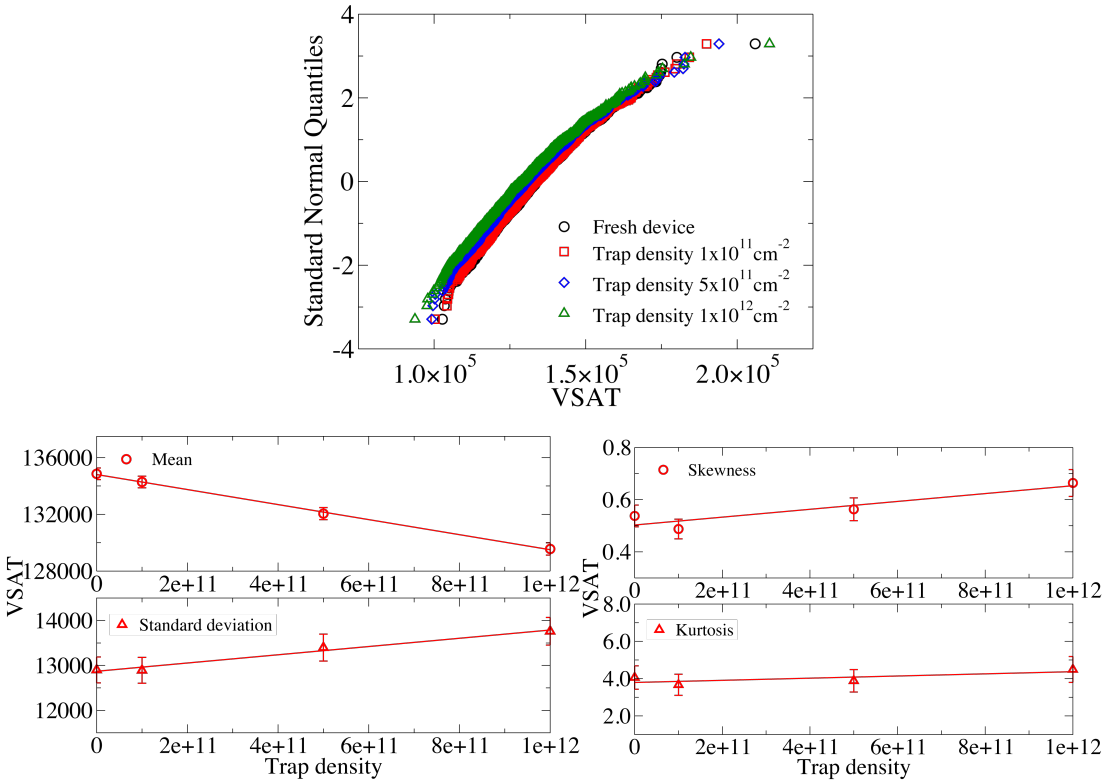


Fig.4.35 The distribution of VSAT for NMOS.

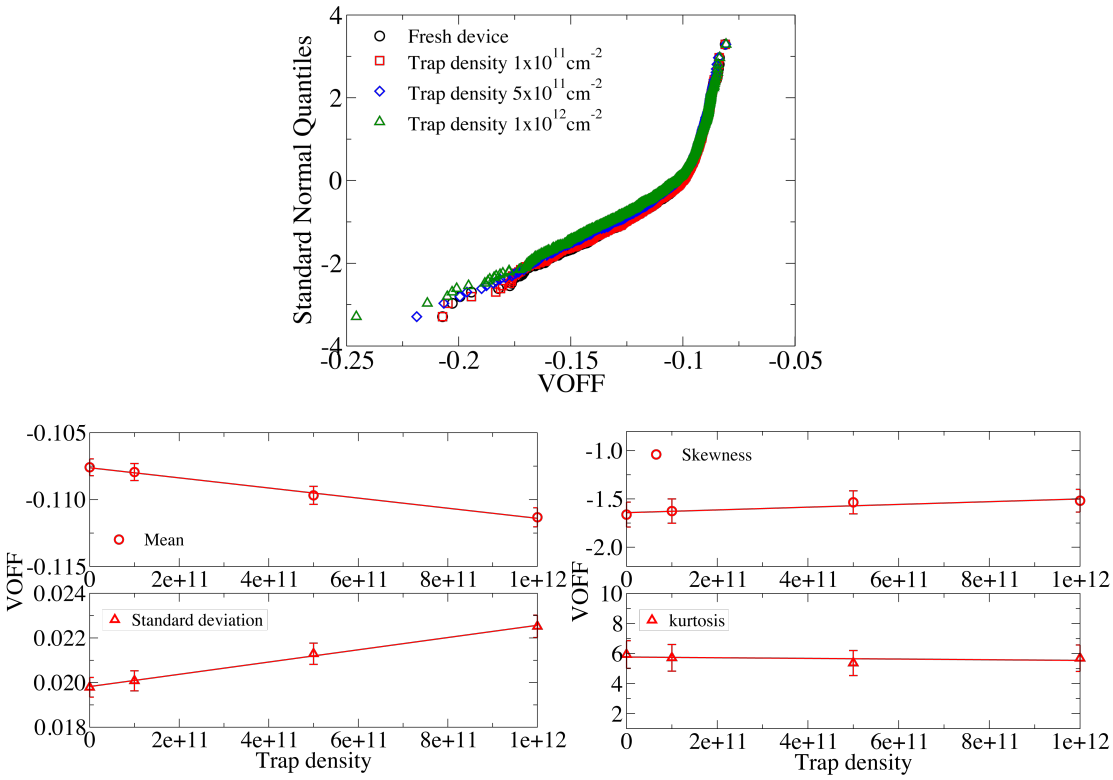


Fig.4.36 The distribution of VOFF for NMOS.

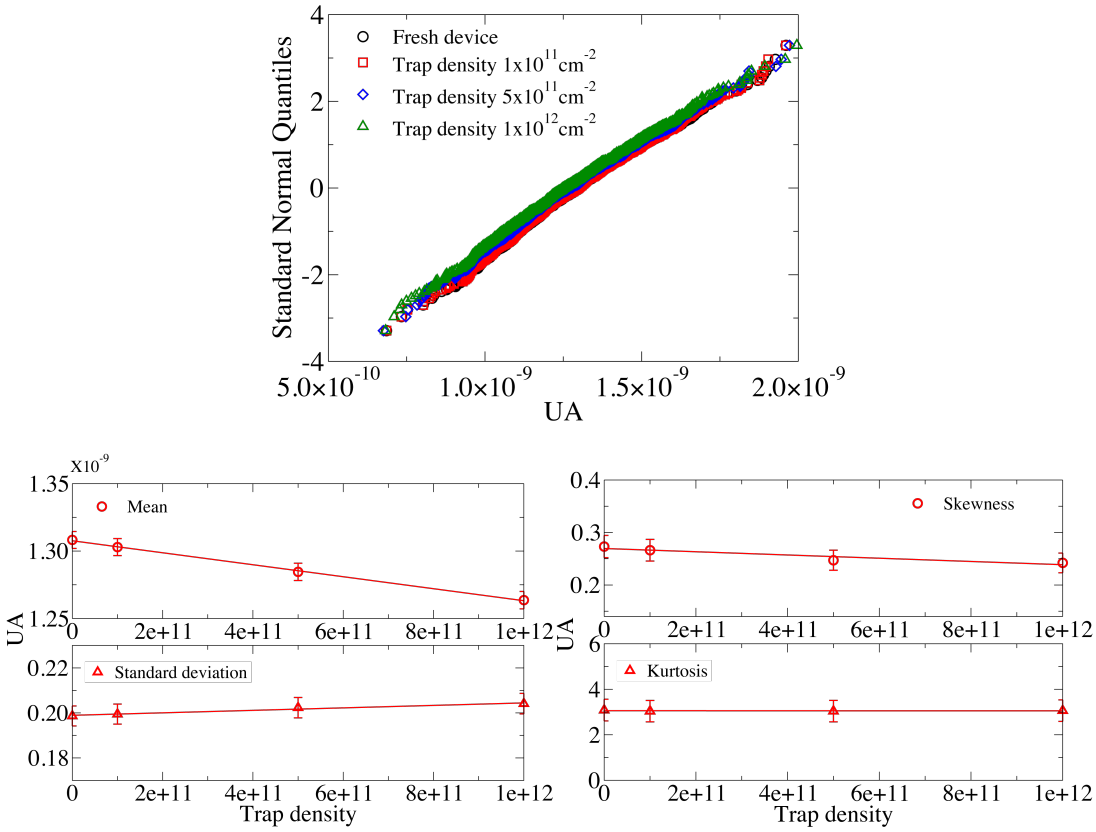


Fig.4.37 The distribution of UA for NMOS.

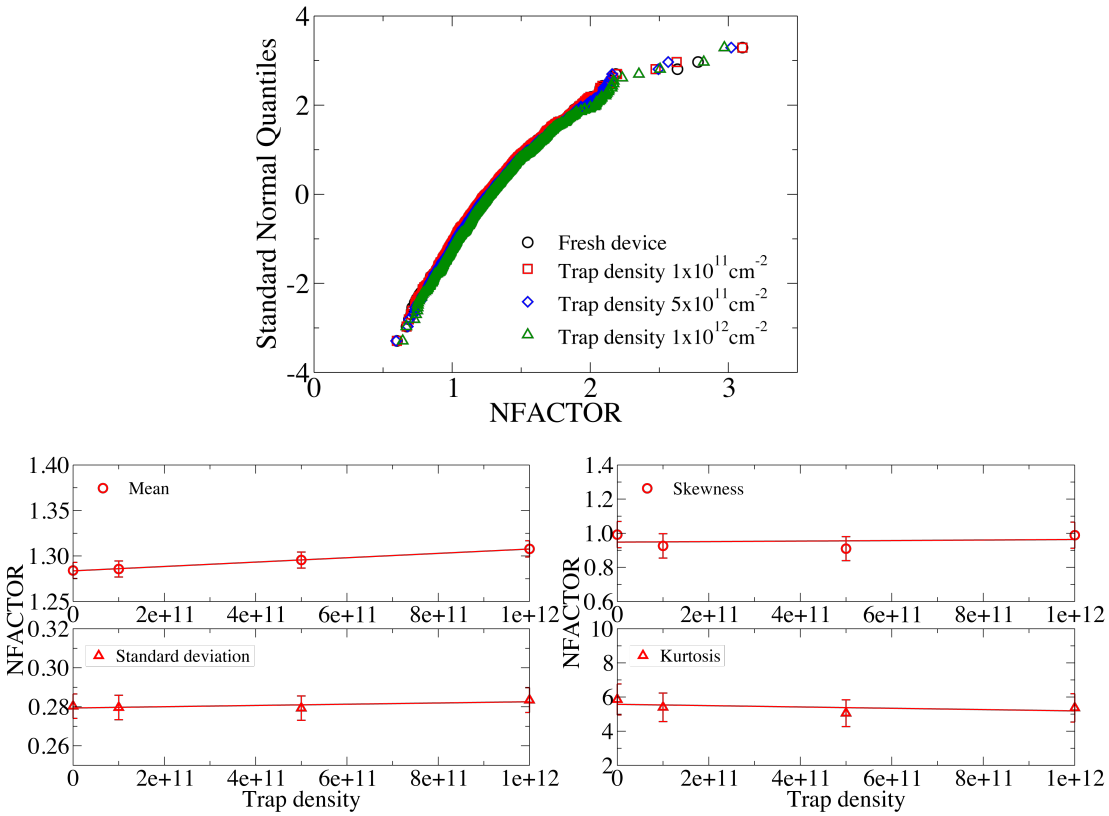


Fig.4.38 The distribution of NFACTOR for NMOS.

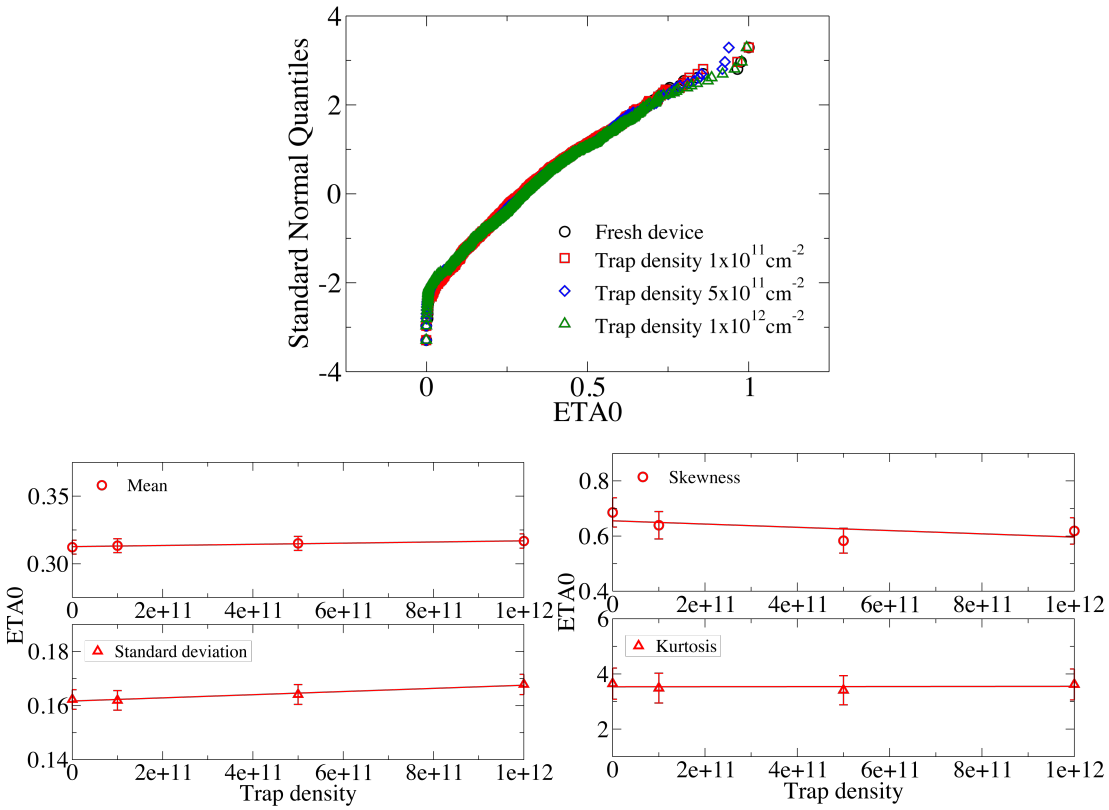


Fig.4.39 The distribution of ETA0 for NMOS.

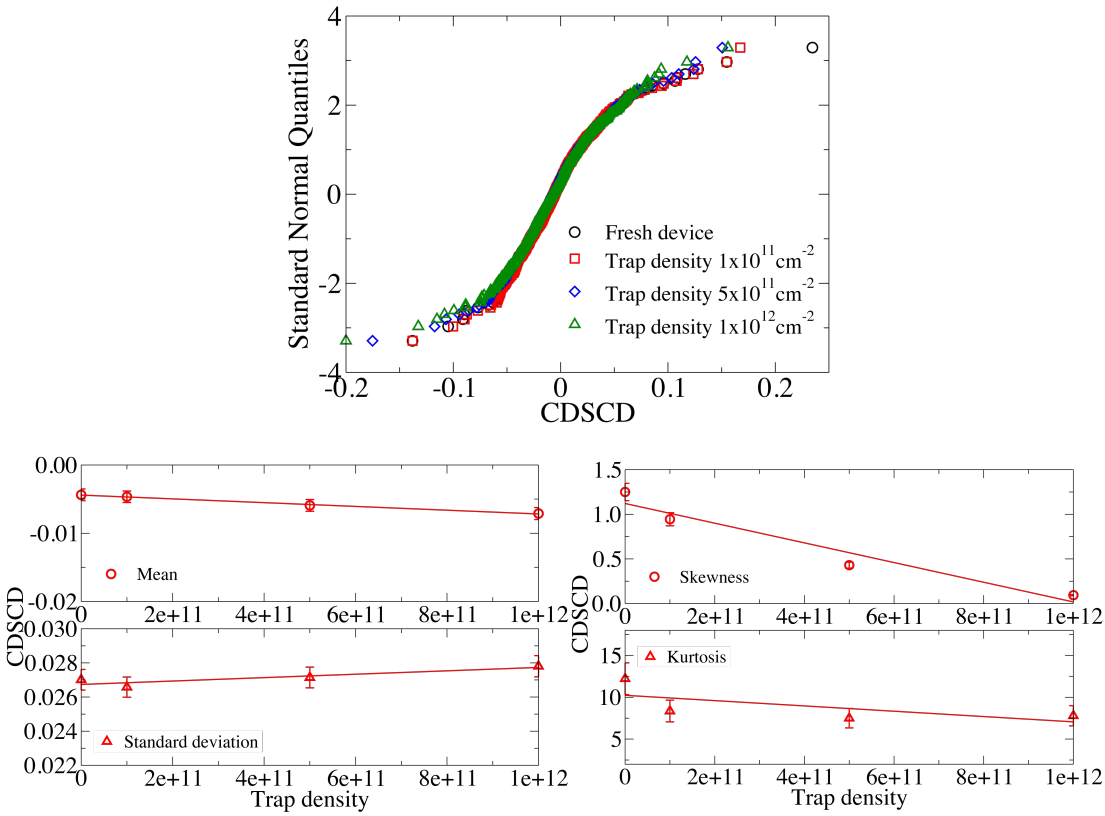


Fig.4.40 The distribution of CDSCD for NMOS.

4.5 Summary

In this chapter, resource data that represent the 25 nm physical gate length MOSFETs' performance under the influence of statistical variability and BTI-induced ageing are obtained by physical simulations and analysed for a better understanding of device performance. By applying the two-stage compact model extraction strategy, the 25 nm compact models are successfully extracted from the resource data. The extracted compact models can not only capture the figures of merit, but also successfully capture the correlations between these figures of merit consistent with the original physical simulation results. This verifies the accuracy of the statistical compact model extraction. The seven re-extracted parameters in the second stage are also analysed for the preparation of the compact model generation in Chapter 5.

Although the directly extracted compact models can accurately represent physical simulation results, they are only 1,000 of them – one for each device simulated in the statistical device ensemble. They cannot be applied directly to large scale statistical

circuit simulation as the results will suffer from artefacts associated with subsampling and will be unable to make accurate predictions for extremely rare occurrences in the tails of the output distribution. In addition, the trap densities associated with ageing are fixed at specific values representing discontinuous increases in ageing. These problems will be addressed in the generation of statistical compact model that will be illustrated in details in Chapter 5.

Chapter 5

Compact Model Generator

5.1 Introduction

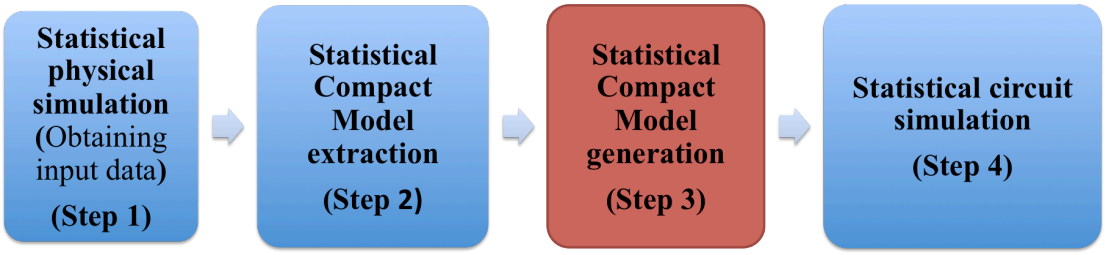


Fig.5.1 Step 3 of the simulation flow

In Chapter 4, statistical compact models were successfully extracted to represent the characteristics for each simulated device within the statistical ensemble. However, although each extracted model can accurately reproduce the performance of each TCAD simulated device, the number of compact models is pre-defined by the number of devices in the physical simulation ensemble. The finite number of input models severely limits the usefulness of the method for the investigation of circuit performance subject to BTI at high sigma, due to subsampling issues.

In addition, device ageing is a continuous process. Devices under different accumulated active time and different stress conditions are aged at different levels. Physical simulations have been carried out at four separate ageing levels (fresh, low, medium, high). The limited number of available levels of ageing limits the accuracy to which BTI-

induced ageing in circuits may be investigated. Compact models at arbitrary ageing level are necessary in order to represent devices at any stage of the continuous ageing process, from newly produced to the end of life.

Since the key physical mechanism behind ageing is trap generation at the gate oxide, trap density is used directly in place of physical age. However, it is more convenient for the circuit designer to input the stress time rather than the trap density in circuit simulation, so an ageing model that can translate the stress time into the trap density is needed.

In this chapter, we address the compact model generation methodology (shown in the red block of Fig.5.1). With this method, a sufficiently large number of compact models can be generated, resolving the limitation imposed by the finite number of directly extracted models. The model generation method creates compact model parameters, with which device performance is produced, preserving both the distributions and correlations of figures of merit of the physical simulations. The development of an interpolation methodology enables statistical compact model to be generated at any arbitrary ageing level. The capability to generate such intermediate ageing models at trap densities that were not physically simulated, has important application in statistical circuit simulation, as this will open up the possibility to include reliability assessment in the circuit design. Simultaneously, an ageing model that can transfer trap density to stress/ageing time is integrated to assist circuit design.

In section 5.2, the subsampling problem is discussed. The traditional compact model generation methods that are used in SPICE software are reviewed in section 5.3.1, while the new generation methodology of generating a large compact model ensemble is introduced in section 5.3.2. In section 5.4, the approach based on interpolation is presented, allowing compact models to be generated at arbitrary trap densities between the lowest (0) and highest ($1x10^{12}cm^{-2}$) trap concentration levels. Section 5.5 introduces an ageing model that links trap density to stress/ageing time. In section 5.6, all the above methodologies are incorporated into RandomSpice. In section 5.7, the generated compact models are compared to the reference data obtained by physical simulations, validating the accuracy of the compact model generation methodologies. Section 5.8 summarises this chapter.

5.2 Subsampling Problem.

As stated above, the finite number of the directly extracted compact models (in our case 1,000) will result in subsampling issues if they are directly used in the circuit simulation, especially in smaller circuits and where one or two critical transistors determine performance. This is due to the fact that only the 1,000 models extracted can only represent these critical transistors.

In order to clearly illustrate the subsampling problem, an example is provided by using the statistical analysis software R. Numbers are randomly picked from the Gaussian distributed population, with a sample of 100 and 200 respectively for 1,000 times. This emulates running 1,000 circuits with different sizes of input models. The distributions are shown in Fig.5.2. The black circles are the numbers picked from the population. They are continuous and smooth and emulate the real circuit condition. Compared with the black circles, the red squares from 200 samples begin to show truncation in the distribution at 2 sigma. The truncations and discontinuities are seen to be more serious for the distribution of the green triangles, which is generated from the sample size of 100. It is obvious that both 100 and 200 input models do not contain sufficient information to represent the underlying real population, especially the rare points that occur at high sigma. The lack of this information directly leads to the truncation and discontinuity at the tails. However, it is the tails of the distribution that are critically important for accurately determining, for example, circuit yield, as this demands an understanding of distributions at high sigma, typically 6 sigma and above. This subsampling problem therefore greatly limits high sigma investigation.

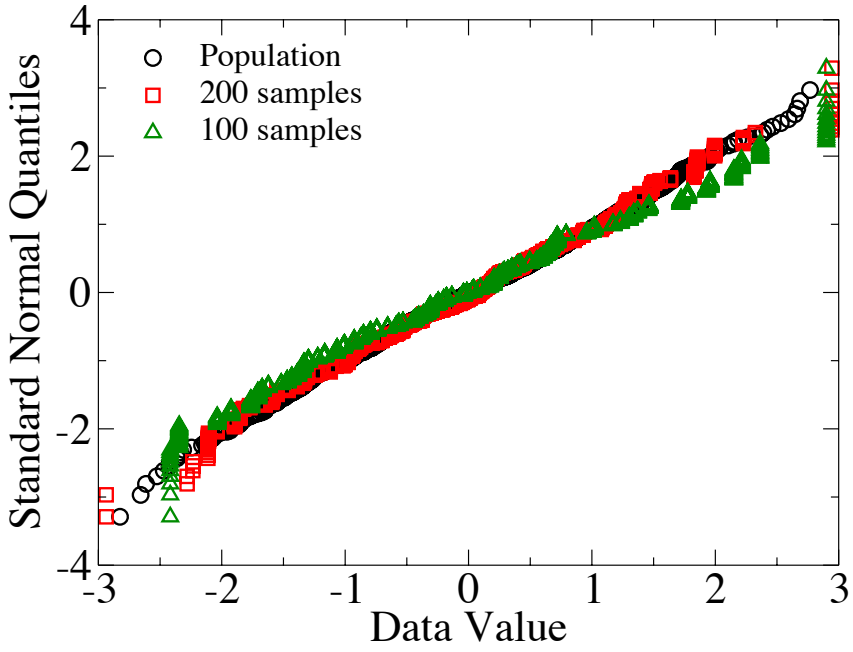


Fig.5.2 Subsampling issue.

In order to accurately simulate the circuit performance, an ensemble of billions, or even larger, compact models is needed in the library to ensure the distribution of circuit performance is accurately captured, specifically including the tails. However, the direct extraction method becomes impractical for such a large number of models because of the time and computational effort required to generate the models from TCAD simulation, as well as the impracticality of working with the necessarily large database that is needed to store and manipulate the extracted look-up table models. A method of generating sufficiently large and accurate compact models is necessary. Such generation methodologies exist, and we will investigate them in this chapter.

5.3 Compact Model Generation Methodology.

5.3.1 Gaussian V_T

There are some existing approaches of generating statistical compact models, such as the Gaussian V_T method, the extension of Gaussian V_T method and the method that is based

on Principal Component Analysis (PCA) [89, 90, 97] etc. The assumption of a simple Gaussian V_T distribution is a popular method implemented in many commercial SPICE simulators for statistical circuit simulations. Since V_{TH} is the first order effect of statistical variability and ageing, and the parameter VTH0 in BSIM4 model accounts for low drain V_{TH}, a common method is to generate VTH0 on the fly while keeping other parameters the same to produce the large ensemble of compact models. The VTH0 generation is based on the assumption of Gaussian distributions for VTH0 by using the mean and standard deviation from the directly extracted models (or by measuring V_{TH} from the underlying technology and calculating the corresponding mean and standard deviation). Gaussian V_T method is popular due to the simplicity and speed. It is because 1) if measuring from the device, only one measurement of V_{TH} is needed to be taken for each device. 2) In addition, only one parameter VTH0 needs to be generated based on the first two moments. Gaussian parameter generation is numerically quick and efficient.

Fig.5.3 shows the comparison of the distribution of figures of merit between physical simulation results and generated compact models using the simple Gaussian V_T method for fresh NMOS devices. It can be seen that the Gaussian V_T generation method can capture the V_{TH} distribution at low drain bias since this method is based on generation VTH0 which account for low drain V_{TH}. However, it does not capture well V_{TH} at high drain bias. This is mainly because some other parameters that vary the subthreshold slope are not considered, especially ETA0, which accounts for DIBL and can vary the high drain V_{TH}, is a constant in this case. For I_{ON} and I_{OFF}, it is clear that both at low drain and high drain bias the distributions have some mismatches against the physical simulation results, and the distributions at high drain bias have a larger mismatch than at the low drain bias. These distributions are not totally lost but are not accurately captured is due to the complex correlations of the compact model parameters. VTH0 is correlated with the I_{ON} and I_{OFF}, however, the parameters such as VSAT and VOFF are not considered in this approach but they vary on-current and off-current. In this method, the distribution of DIBL is totally lost because the value of ETA0, which mainly captures DIBL effect, is kept constant.

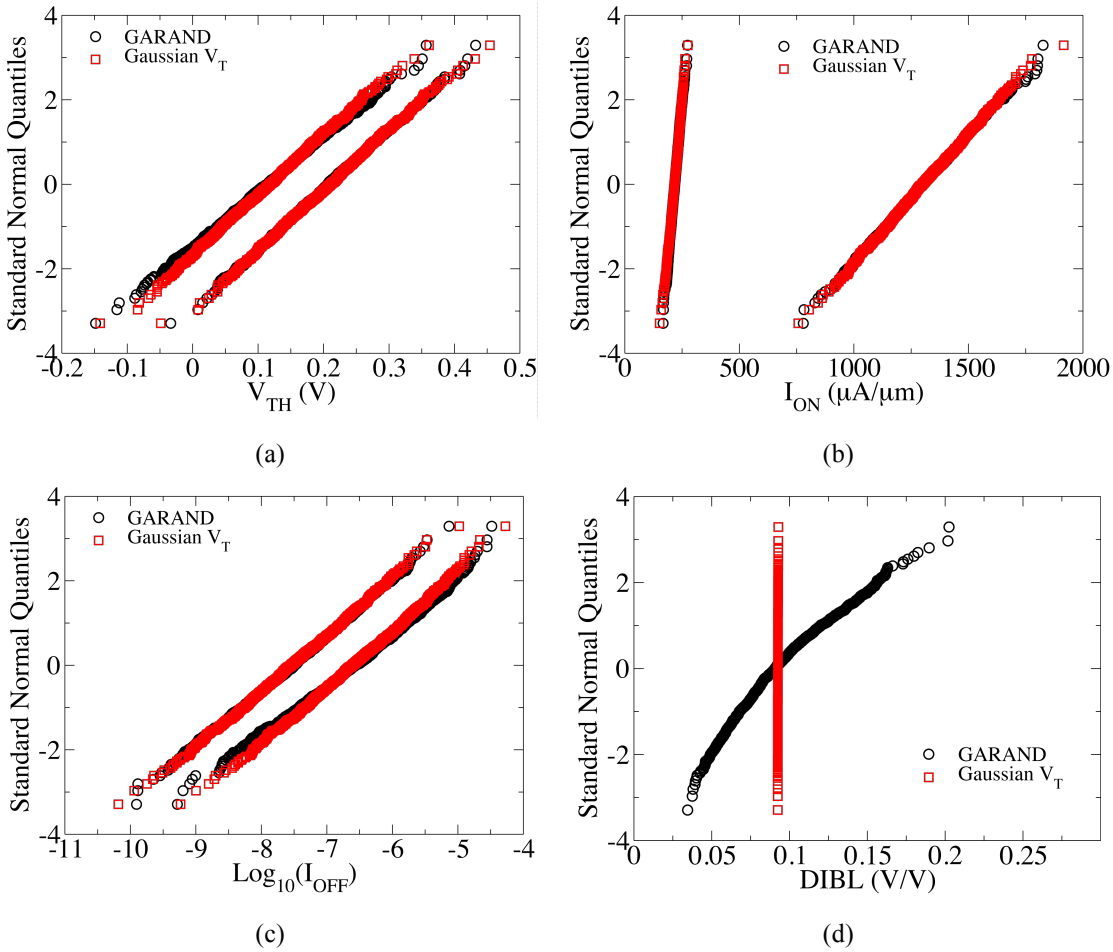


Fig.5.3 The comparisons of (a) V_{TH}, (b) I_{ON}, (c) $Log_{10}(I_{OFF})$, (d) DIBL for fresh NMOS devices, between physical simulation results and Gaussian V_T generated compact models at high drain and low drain bias.

Fig.5.4 and Fig.5.5 show the correlations of figures of merit from physical simulations (shown in black) and Gaussian V_T generated compact models (shown in red) for a fresh NMOS device. The correlations show the complex physical relationships between figures of merit in a single figure. These figures show that Gaussian V_T generated compact models give completely incorrect information on the correlations (as a single parameter generation assumes everything is one-to-one correlated with threshold voltage). Figures of merit resulting from assuming a simple Gaussian V_T approach, are entirely correlated with each other and are governed by a shift in only one parameter, VTH0.

Although the Gaussian V_T approach is simple and easy, from the analysis above, this method cannot be used to represent physical simulations accurately. It is because the simple assumption regarding VTH0 distribution. The usage of one parameter to reflect

all effects is obviously insufficient, since other physical parameters that cause additional variations are missing.

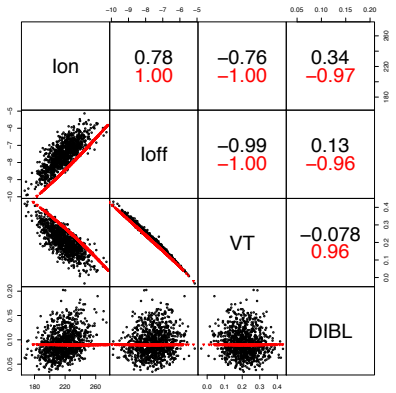


Fig.5.4 Correlations between figures of merit for fresh NMOS devices (V_{DS}=0.05V). The black indicates physical simulation results. The red indicates Gaussian V_T generated compact model results.

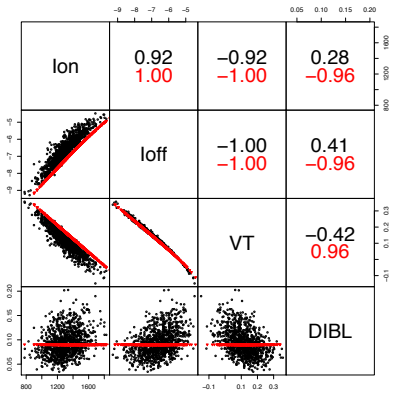


Fig.5.5 Correlations between figures of merit for fresh NMOS devices (V_{DS}=1V). The black indicates physical simulation results. The red indicates Gaussian V_T generated compact model results.

Rather than using one parameter, other approaches, such as the extension of Gaussian V_T method or PCA etc. generate several key parameters to produce the new compact models. However, all these approaches are based on the assumption that the selected parameters' distributions are Gaussian, which contradicts the fact that some parameters have highly non-Gaussian distribution. We can take as an example of the VOFF distributions in section 4.3.2 from the extracted compact models. Therefore, these approaches introduce errors and result in optimistic or pessimistic conclusions in circuit simulations [97]. In section 5.3.2, the new generation methodology will be introduced, with an assumption of a wide range of distribution shapes.

5.3.2 Generalized Lambda Distribution

GLD is based on the Tukey Lambda distribution [98]. It is based on four-parameter generalization ($\lambda_1, \lambda_2, \lambda_3$ and λ_4) and there are two generalizations for GLD, the traditional RS [99] and the FMKL [100] parameterization. The RS parameterization is proposed by Ramberg and Schmeiser in 1974, however, certain combinations of parameters lead to invalid probability distribution using RS while this does not occur in

the FMKL parameterization. Therefore, in this research, the FMKL based GLD is used. A key feature of the FMKL approach is the inherent flexibility as it can fit a wide range of the distribution shapes [101]. This approach's simplicity is manifested in obtaining the quantile function (the inverse of the Cumulative Distribution Function (CDF)) by using the first four moments of the sample data. Therefore, the compact model parameter's quantile function and CDF can be obtained by calculating the first four moments of the statistically extracted compact model parameters. Then a large ensemble of compact model parameters can be generated based on the randomly generated CDF numbers that are used as the input into each parameter's quantile function. These compact model parameters can be inserted into the uniform model to form infinite compact models.

In the FMKL based GLD, four parameters ($\lambda_1, \lambda_2, \lambda_3$ and λ_4) constitute the quantile function. The quantile function is defined as:

$$Q(u) = \lambda_1 + \frac{1}{\lambda_2}\left(\frac{u^{\lambda_3}-1}{\lambda_3} - \frac{(1-u)^{\lambda_4}-1}{\lambda_4}\right) \tag{5.1}$$

$\lambda_1, \lambda_2, \lambda_3,$ and λ_4 can be represented by the complex equations using the first four moments from the sample. Therefore, the values of λ_1, λ_2, λ_3, and λ_4 can be fitted if the moments are given. The corresponding CDF can be derived using the quantile function with the values of λ_1, λ_2, λ_3, and λ_4 obtained from the moments. The details are shown in [101].

In the second stage of extraction, seven parameters of VTH0, VSAT, VOFF, UA, NFACTOR, EAT0 and CDSCD are re-extracted at each ageing level. By calculating the first four moments from these re-extracted parameters and using them to find λ_1, λ_2, λ_3, and λ_4, compact model parameters' quantile function and distribution function can be obtained at each ageing level. Thus, random CDF numbers are required as the input for the quantile function. Then compact model parameter values can be generated following the distribution.

In order to have CDF numbers, columns of random numbers are generated using multivariate normal distribution method based on the correlation matrix. The correlation matrix ensures the generated numbers between each column preserve the correlations between the generated compact model parameters. This will be discussed in the next

paragraph. In this study, seven columns of random numbers are generated because seven parameters are used. Then, each column of these numbers are applied using the Probability Integral Transform on the Gaussian Cumulative Distribution Function (CDF) and a random sample of uniform distributed CDF numbers belonging to U(0,1) are obtained. Using these numbers to generate the new compact model parameters, and inserting each set of the generated compact model parameters into the uniform compact model, a sufficiently large number of compact models can be generated.

It should be noted that the correlation coefficients in the correlation matrix are not the correlations calculated from the re-extracted parameters. This is because correlations are changed after the Probability Integral Transform. The matrix of correlations between CDF numbers of each corresponding parameter are obtained using numerical root finding, using the compact model parameter correlations as the initial guess.

Fig.5.6 to Fig.5.12 show the new generated parameter values against the extracted compact model parameter values at trap density of 0 and $1x10^{12}cm^{-2}$ respectively. From these figures, it is clear that the GLD method can accurately re-generate the new parameter values by calculating the first four moments from the re-extracted parameters. Here, it is shown in Fig.5.8, even if the original distribution of parameter VOFF is very pathological and difficult to model, the results still show a good agreement to some degree.

Fig.5.13 and Fig.5.14 show the scatter plots and correlations between the seven parameters at trap density of 0 and $1x10^{12}cm^{-2}$ respectively. The black is the results from extracted compact models and the red is the results from generated compact models using GLD method. It shows that the generated parameters using GLD highly agree with the extracted parameter correlation results, validating the accuracy of correlation transform method stated above.

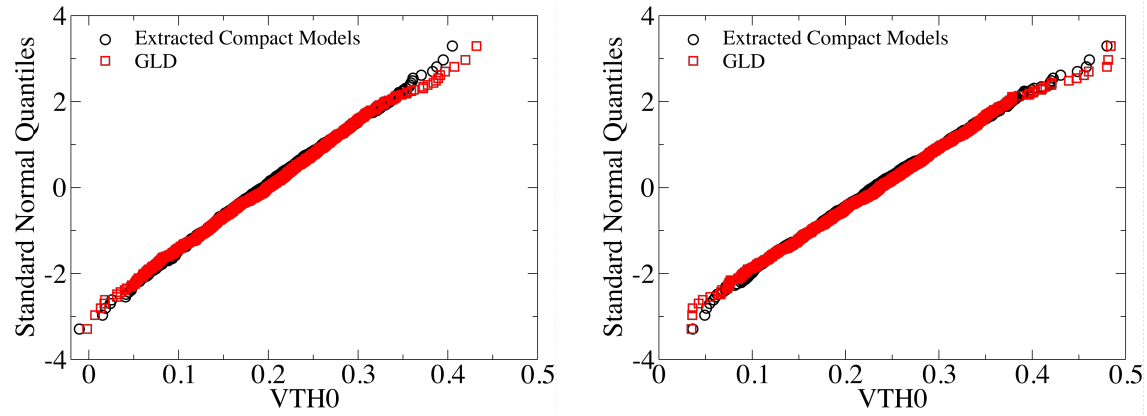


Fig.5.6 The comparisons of VTH0 values between the extracted compact models and regeneration by GLD for NMOS. The left plot is at trap density of 0 while the right plot is at trap density of $1x10^{12}cm^{-2}$.

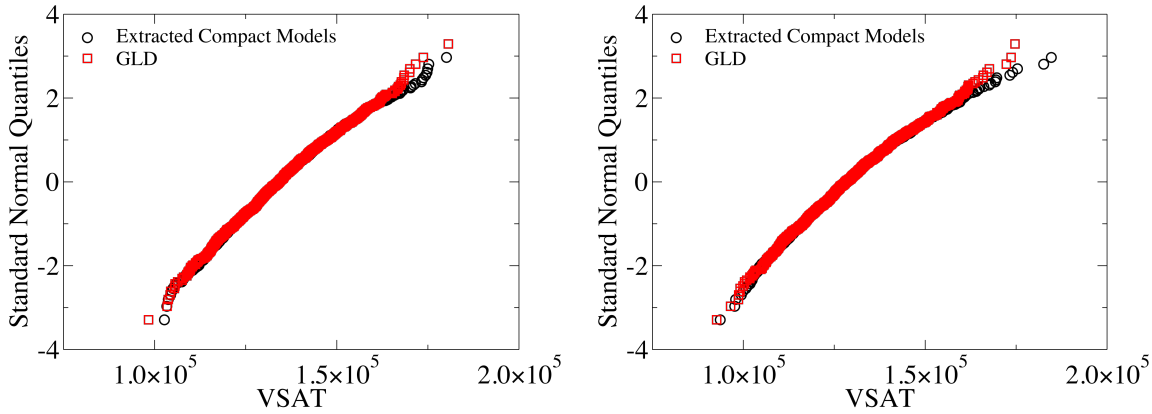


Fig.5.7 The comparisons of VSAT values between the extracted compact models and regeneration by GLD for NMOS. The left plot is at trap density of 0 while the right plot is at trap density of $1x10^{12}cm^{-2}$.

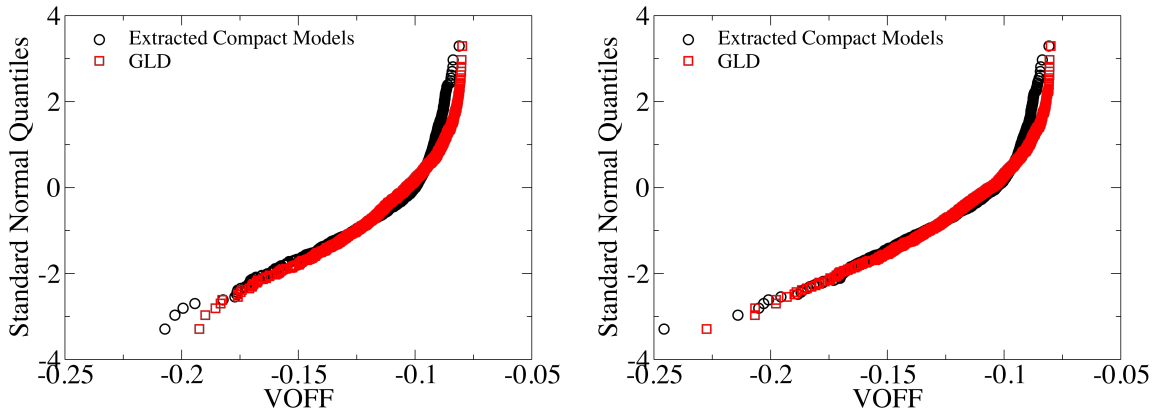


Fig.5.8 The comparisons of VOFF values between the extracted compact models and regeneration by GLD for NMOS. The left plot is at trap density of 0 while the right plot is at trap density of $1x10^{12}cm^{-2}$.

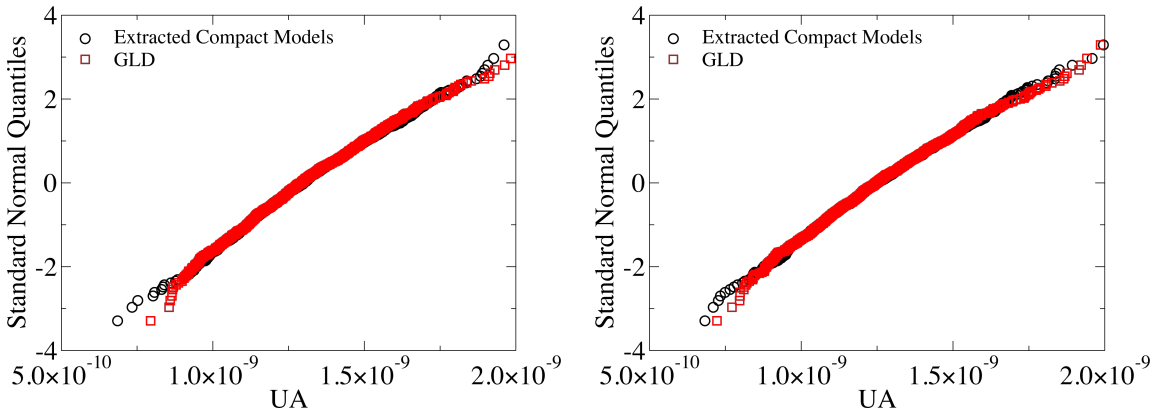


Fig.5.9 The comparisons of UA values between the extracted compact models and regeneration by GLD for NMOS. The left plot is at trap density of 0 while the right plot is at trap density of $1x10^{12}cm^{-2}$.

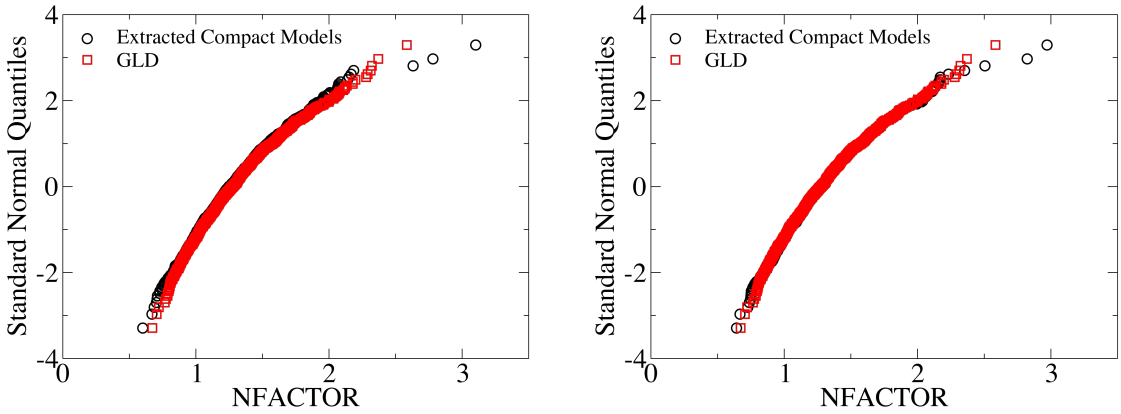


Fig.5.10 The comparisons of NFACTOR values between the extracted compact models and regeneration by GLD for NMOS. The left plot is at trap density of 0 while the right plot is at trap density of $1x10^{12}cm^{-2}$.

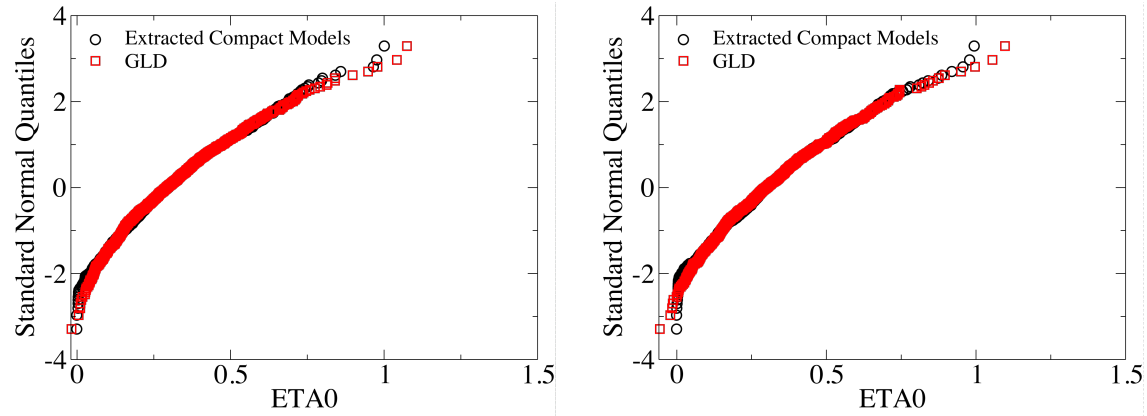


Fig.5.11 The comparisons of ETA0 values between the extracted compact models and regeneration by GLD for NMOS. The left plot is at trap density of 0 while the right plot is at trap density of $1x10^{12}cm^{-2}$.

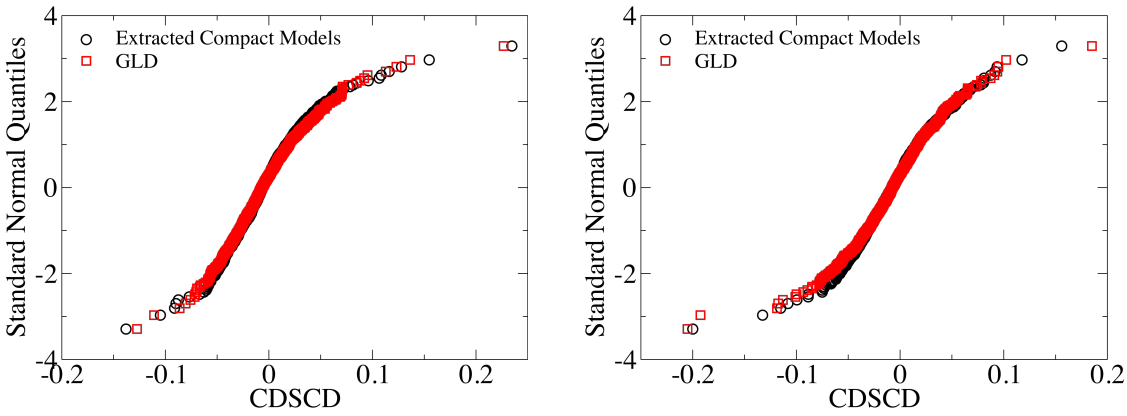


Fig.5.12 The comparisons of CDSCD values between the extracted compact models and regeneration by GLD for NMOS. The left plot is at trap density of 0 while the right plot is at trap density of $1x10^{12}cm^{-2}$.

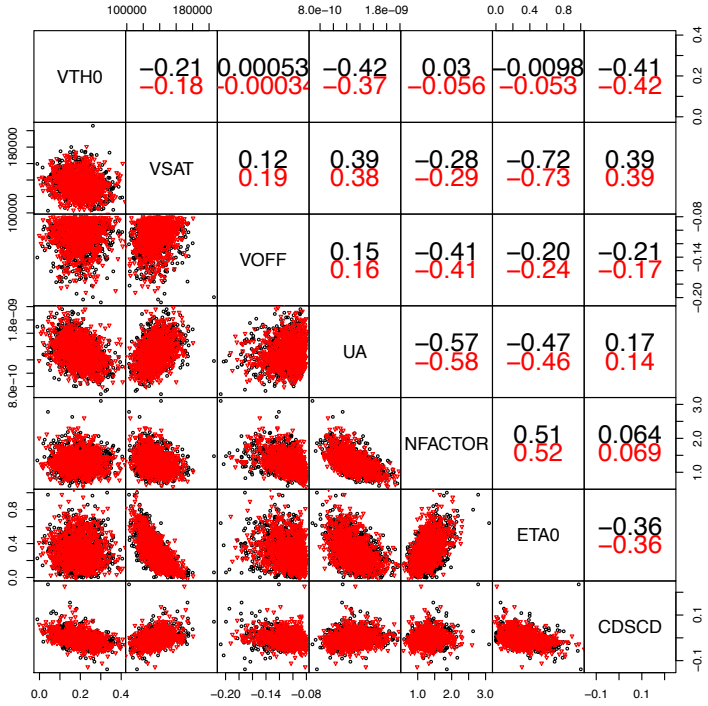


Fig.5.13 The scatter plots and correlations between the seven parameters at trap density of 0 for NMOS. The black is the results from extracted compact models, the red is the results using GLD.

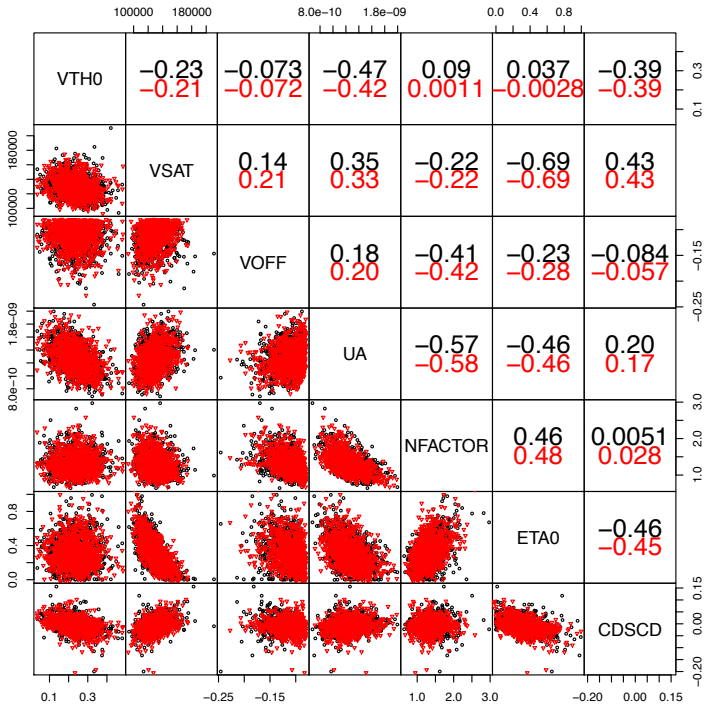


Fig.5.14 The scatter plots and correlations between the seven parameters at trap density of $1x10^{12}cm^{-2}$ for NMOS. The black is the results from extracted compact models, the red is the results using GLD.

The GLD approach is incorporated into RandomSpice using the patented ModelGen technology. The contribution of this chapter is not the development of the ModelGen technology, patented by GSS, but the utilization on demonstration of the technology in a real life example. Using RandomSpice, 1,000 compact models are generated with this approach for each ageing level. Fig.5.15 to Fig.5.18 show the comparisons of figures of merit between physical simulations and the generated compact models at trap densities of 0 and $1x10^{12}cm^{-2}$ respectively, while Fig.5.19 to Fig.5.22 show the correlation comparisons at each ageing level. These figures verify the accuracy of the GLD generation approach, showing that the generated compact models not only highly agree with the distribution of the physical simulation, but can capture the correlations between figures of merit, no matter with statistical variability only or with BTI-induced ageing as well. The corresponding PMOS comparisons are shown in appendix C and the same conclusion can be drawn.

Fig.5.23 shows V_{TH} comparison between the 10,000 generated devices and the 1,000 simulated devices at trap density of $1x10^{12}cm^{-2}$ at high drain bias. It shows that the

generated devices follows the distribution of the simulated devices well and accurately extends the trends. This verifies the accuracy of this generation method for high sigma investigation.

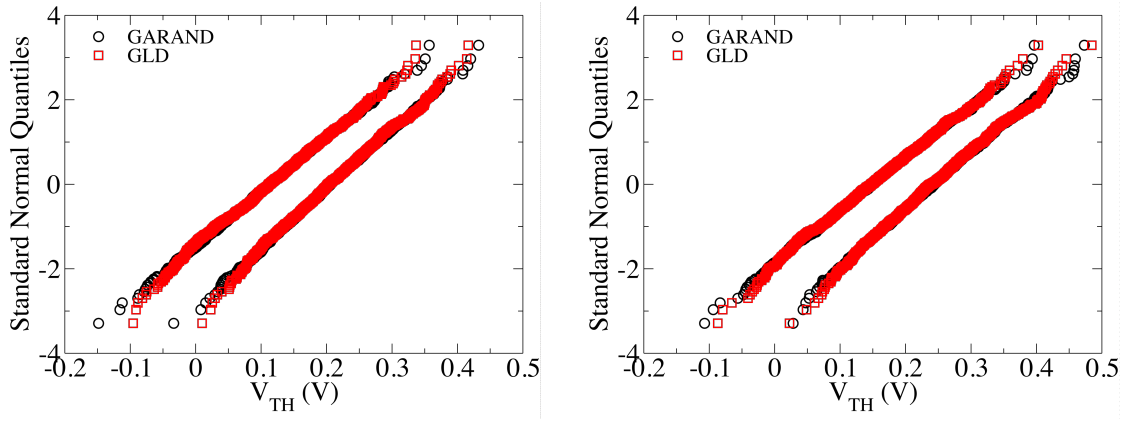


Fig.5.15 Comparisons of V_{TH} between physical simulation and GLD generated compact models for NMOS devices. The left and right figures are at trap density of 0 and $1x10^{12}cm^{-2}$ respectively.

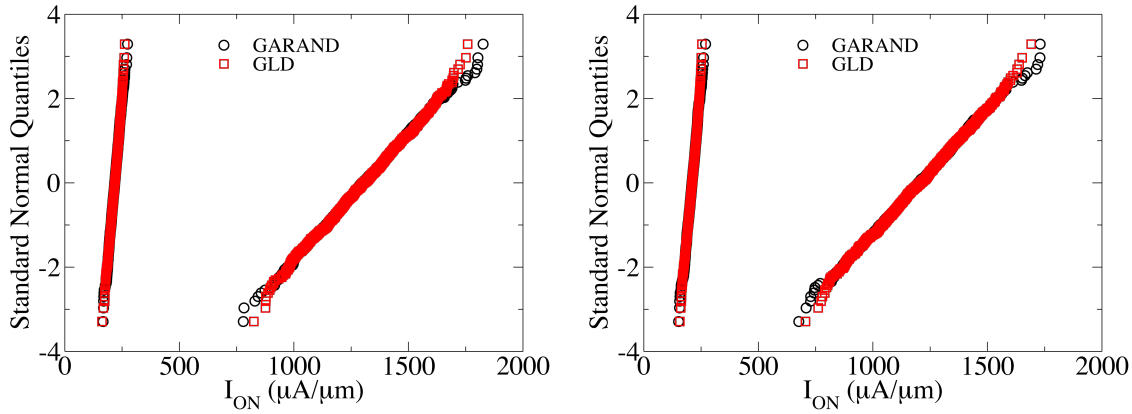


Fig.5.16 Comparisons of I_{ON} between physical simulations and GLD generated compact models for NMOS devices. The left and right figures are at trap density of 0 and $1x10^{12}cm^{-2}$ respectively.

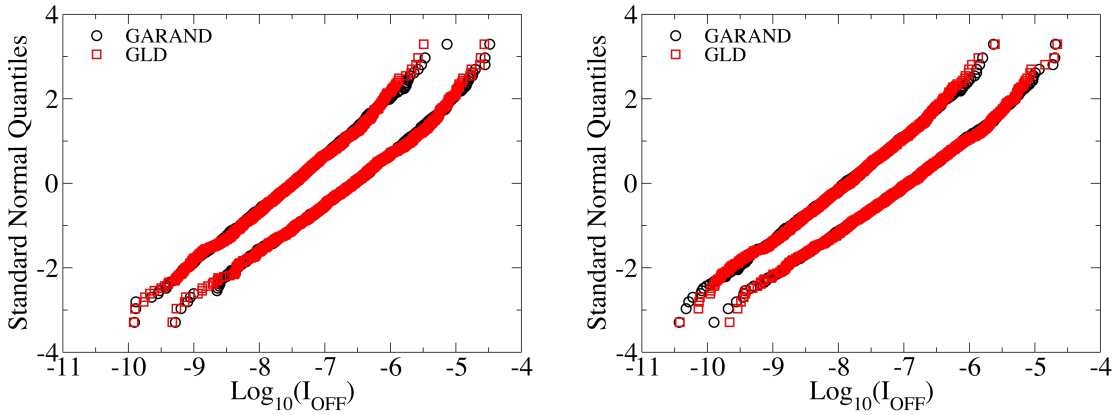


Fig.5.17 Comparisons of I_{OFF} between physical simulations and GLD generated compact models for NMOS devices. The left and right figures are at trap density of 0 and $1x10^{12}cm^{-2}$ respectively.

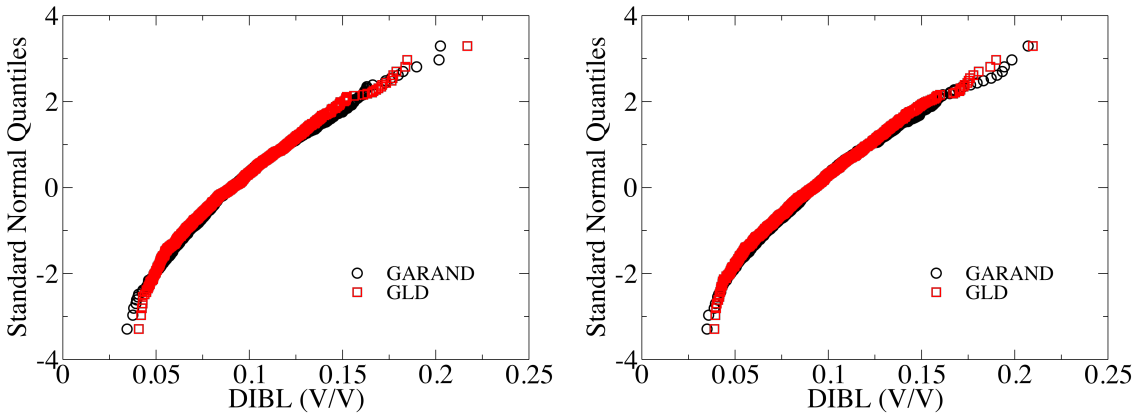


Fig.5.18 Comparisons of DIBL between physical simulations and GLD generated compact models for NMOS devices. The left and right figures are at trap density of 0 and $1x10^{12}cm^{-2}$ respectively.

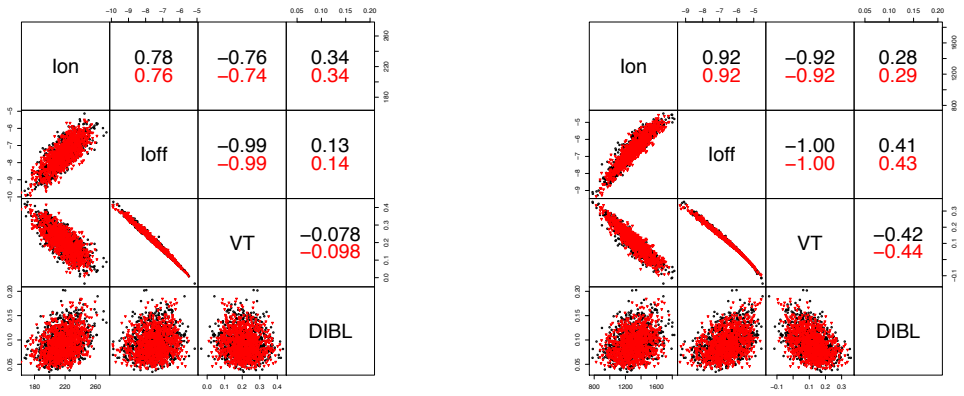


Fig.5.19 Correlations of figures of merit of NMOS between physical simulation and GLD generated compact models at trap density of 0. The left figure is when V_{DS}=0.05V, while the right figure is when V_{DS}=1V.

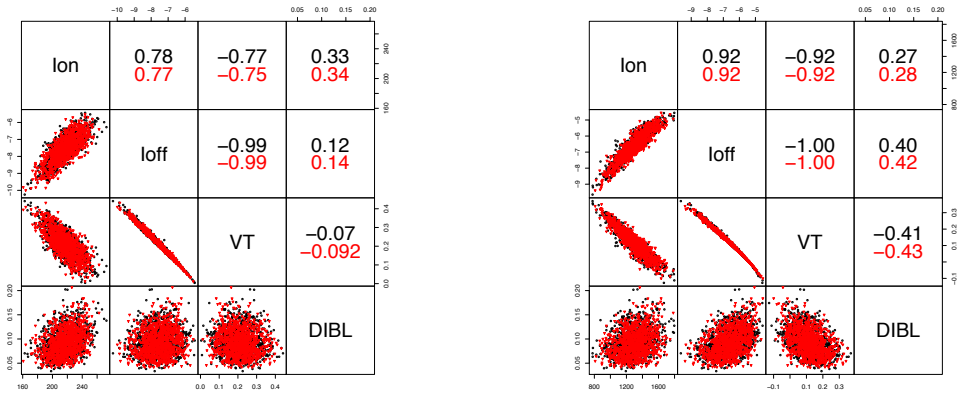


Fig.5.20 Correlations of figures of merit of NMOS between physical simulation and GLD generated compact models at trap density of $1x10^{11}cm^{-2}$. The left figure is when V_{DS}=0.05V, while the right figure is when V_{DS}=1V.

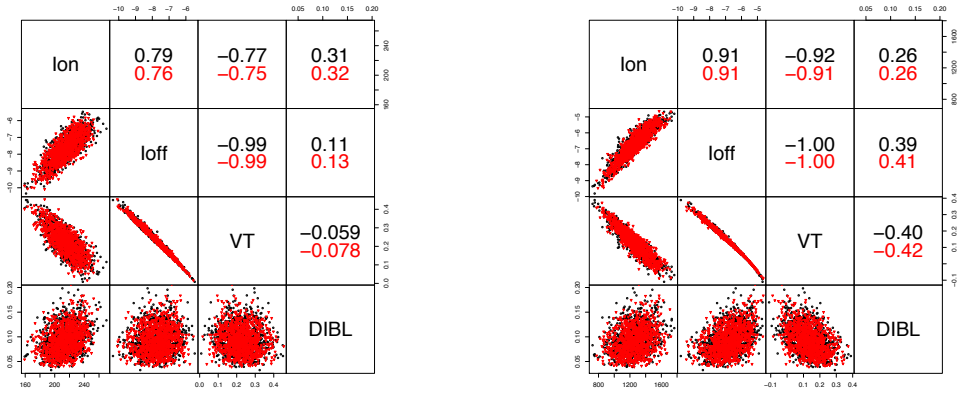


Fig.5.21 Correlations of figures of merit of NMOS between physical simulation and GLD generated compact models at trap density of $5x10^{11}cm^{-2}$. The left figure is when V_{DS}=0.05V, while the right figure is when V_{DS}=1V.

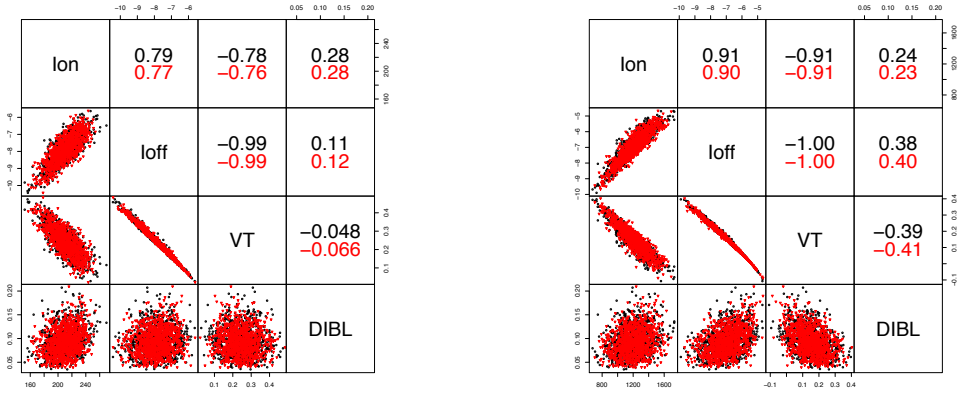


Fig.5.22 Correlations of figures of merit of NMOS devices between physical simulation and GLD generated compact models at trap density of $1x10^{12}cm^{-2}$. The left figure is when V_{DS}=0.05V, while the right figure is when V_{DS}=1V.

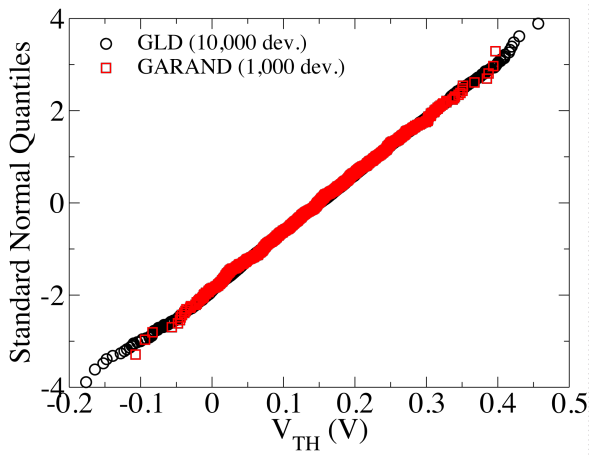


Fig.5.23 Comparisons of V_{TH} between physical simulation and GLD generated compact models for NMOS devices at trap density of $1x10^{12}cm^{-2}$.

5.4 Interpolation Between Trap Densities

Using the GLD method, an effectively large number of compact models can be generated accurately based on the extracted seven key compact model parameters. However, these compact models can only be generated at particular ageing levels, which are pre-specified in the physical simulation scenarios (fresh, low, medium and high). In this section, the patented interpolation method is introduced that enables compact models to be generated at any arbitrary ageing level.

To illustrate the approach used in this work that can generate compact models at arbitrary ageing levels, it is important to recall the analysis of the extracted seven key parameters in section 4.3.2. The mean, standard deviation, skewness and kurtosis vary monotonically and linearly with trap density, from which the relationship between moments and trap density for every parameter can be written in the format of $y = ax + b$, where y is the value of a moment for each parameter and x is the trap density. As the moment value at trap density of 0, $1x10^{11}$, $5x10^{11}$, $1x10^{12}cm^{-2}$ are given, line interpolation can be performed and each parameter's four moments can be obtained at any trap density. Using the moments at any trap density as an input into the GLD method, sufficiently large compact models can be generated at arbitrary trap density.

The interpolation method combined with GLD enables the generation of compact models at arbitrary trap densities. This enables representing devices in the circuit stressed with different accumulated time and at different bias conditions. The validation of this method will be shown in section 5.7 of this chapter.

5.5 Translation Between Ageing Time and Trap Densities.

The physical TCAD simulations for this study are performed at specified trapped charge densities, rather than explicitly stress/ageing time. In general, it is the stress/ageing time, rather than the trapped charge density that is convenient in terms of circuit simulations.

Consequently, we have incorporated a proof of concept ageing model in RandomSpice to map the stress time/age into average charge density [102].

First, we assume a power law describing the dependence between stress time and ΔV_T, as observed experimentally in [103, 104]. Using the results from the physical simulation, the relationship between average ΔV_T and charge density is extracted from the results of the physical simulation, shown in Fig.5.25. The power law coefficients are selected such that the simulated average charge densities ($1x10^{11}$, $5x10^{11}$ and $1x10^{12}cm^{-2}$) span the range of approximately 1 μs to 1 year, as illustrated in Fig.5.24. It should be noted that, for the sake of simplicity, the ΔV_T value corresponding to a particular time is assumed to be the *average* ΔV_T when calculating the trap density.

It should also be noted here that due to different technology and material used in the device, the power law coefficient can vary. There are also some disputes about the NBTI and PBTI difference. As a result, the power law coefficient for NMOS and PMOS may vary as well, according to the measurements. For simplicity, in this study we assume that with the same usage time, NMOS and PMOS generate the same trap quantities. Importantly, this is the illustrative case used in this study but is not the only case that the methodologies can handle. The methodologies are sufficient to handle the situation that NMOS and PMOS are at different trap densities; even for different NMOS or PMOS, they can be at different trap densities.

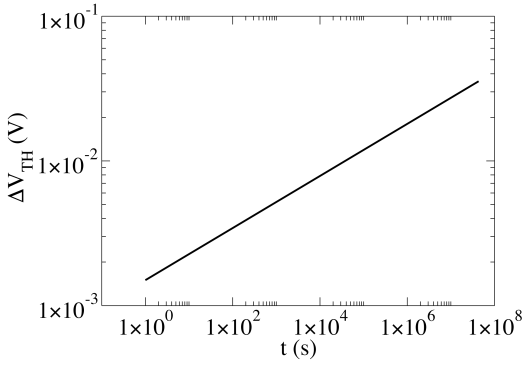


Fig.5.24 Time dependent drift of ΔV_T [102].

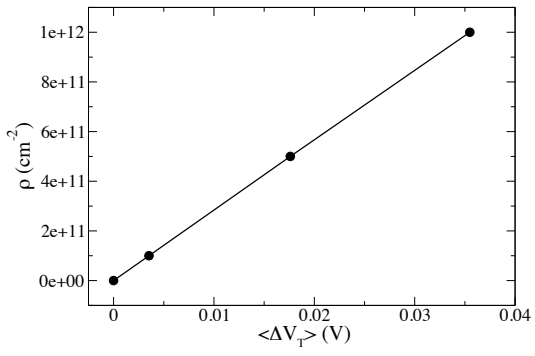


Fig.5.25 Trapped charge density as a function of average ΔV_T [102].

The relations presented in Fig.5.24 and Fig.5.25 have the following form:

$$\Delta V_T(t) = at^b \tag{5.2}$$

$$\rho(\overline{\Delta V_T}) = c + d\overline{\Delta V_T} \tag{5.3}$$

With the time ranging from 1μs to 1year, the equations are as following:

$$\Delta V_T(t) = 1.5\times10^{-3}t^{0.18} \tag{5.4}$$

$$\rho(\overline{\Delta V_T}) = 1\times10^{10} + 2.82\times10^{13}\overline{\Delta V_T} \tag{5.5}$$

With this ageing model, the degradation time corresponding to 0 ("fresh"), $1x10^{11}cm^{-2}$, $5x10^{11}cm^{-2}$ and $1x10^{12}cm^{-2}$ densities is approximately t=0, 66 s, 10 days and 12 months.

5.6 Incorporating into RandomSpice

The GLD, the trap density interpolation method, and the proof of concept ageing model are incorporated into RandomSpice as three blocks. Thus, compact models can be generated at any trap density between 0 to $1x10^{12}cm^{-2}$ using RandomSpice, using the patented ModelGen technology.

Before performing Monte Carlo circuit simulation using RandomSpice, the circuit netlist, input file, compact model library, and a SPICE backend simulator are required. The circuit netlist contains the circuit, options and measurements that will be executed. This circuit netlist is changed from the standard SPICE netlist, with keyword and device ageing information added after the definition of the device that is required to be randomly generated. Fig.5.26 shows a netlist for the illustration. In this netlist, the NMOS transistor M1 will be performed to obtain the I_DV_G curve. The keyword of ATOM:N is added after defining the node connections, this replaces the standard SPICE MOSFET model call. The length and width information and ageing time is added as well. Here, the ageing time of 66.325s is linked to trap density of $1x10^{11}cm^{-2}$ using the ageing model presented in section 5.5. Compact model library is created using RandomSpice and it includes the compact model information with respect to different generation method. For this study, the compact model library is created containing the uniform compact model's parameter value and the seven re-extracted parameter's values at each ageing level (fresh, low, medium and high). As stated in Chapter 3, RandomSpice

is a Monte Carle simulation engine. Therefore, the SPICE backend is required to perform each circuit simulation. In this study, ngSPICE is used. The input file contains simulation settings. For example, the simulation quantities that will be performed, library locations etc., are included in the input file.

```
*IdVg curve for NMOS

M1 DR  GA 0 0    ATOM:N L=25n W=25n AGE=66.325s

VD DR 0 1
VG GA 0 -0.3

.dc VG -0.3 1 0.001

.print dc i(VD)


.OPTIONS NOMOD NOPAGE TEMP=27
.TEMP 27

.END
```

Fig.5.26 The example of the netlist used in RandomSpice.

The randomized compact model generating flow using the GLD method is shown in Fig.5.27. When performing the circuit simulation, RandomSpice identifies the keyword in the netlist. Then it calls the ageing model block to transfer the ageing time to trap density. The corresponding trap density is then used in the GLD block. The GLD block queries the density interpolation block, in which the four moments values of each of the seven parameters are calculated using the information in the library. These moments values are then used in the GLD block. The output is the generated parameters' values, which replace the value in the uniform model and form the new generated compact model used in the circuit. If standard models are used in the netlist, keywords are still required to be added after the device definition in the netlist. The corresponding full set of parameters for the standard compact model are required to be added at the end of the netlist with keywords as the header.

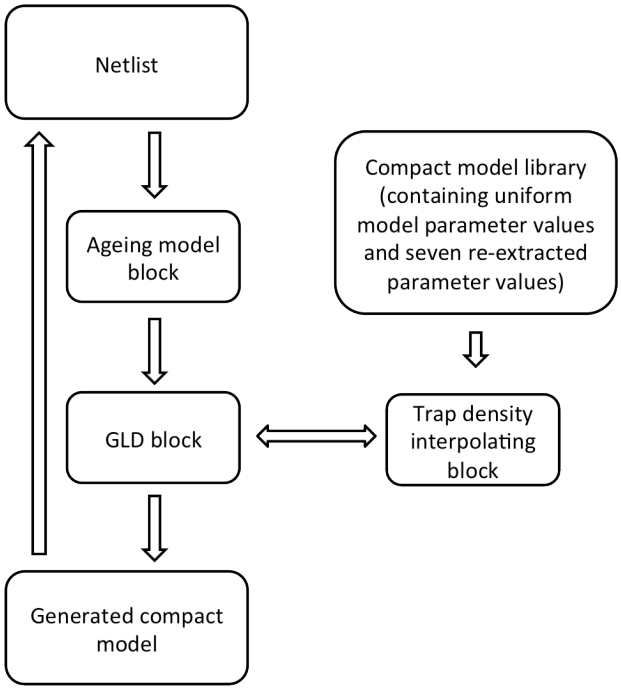


Fig.5.27 Compact model generating flow in RandomSpice.

5.7 Verification of Compact Model Generator

In order to verify the accuracy of the interpolated trap density models, new physical simulations are performed again by GARAND at an intermediate trap density of $7.5x10^{11}cm^{-2}$ (which corresponds to three months' ageing) for validation. Simultaneously, compact models are generated with the ageing time of three months by the generation approach. It is important to emphasize that the TCAD simulated data at t=3 months was not used as an input to RandomSpice's generators. Fig.5.28 to Fig.5.31 compare the distribution of figures of merit and the correlations between them from physical simulation and from compact model generator. These figures are in agreement between the generated models and the physical simulations for both NMOS and PMOS devices, verifying the accuracy of the methods.

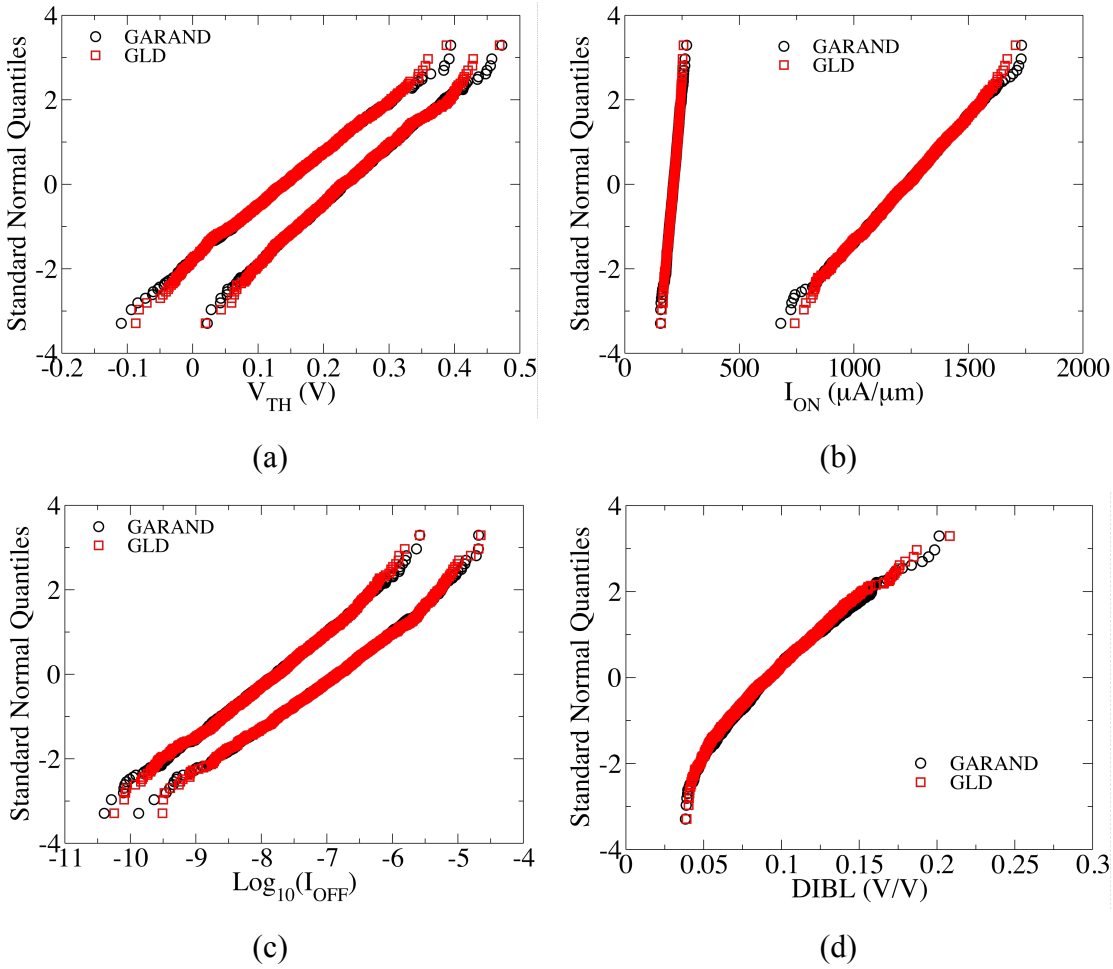


Fig.5.28 Comparisons of figures of merit (a)V_{TH} (b)I_{ON} (c)I_{OFF} (d)DIBL of NMOS between physical simulations and compact model generator at ageing time t=3 month.

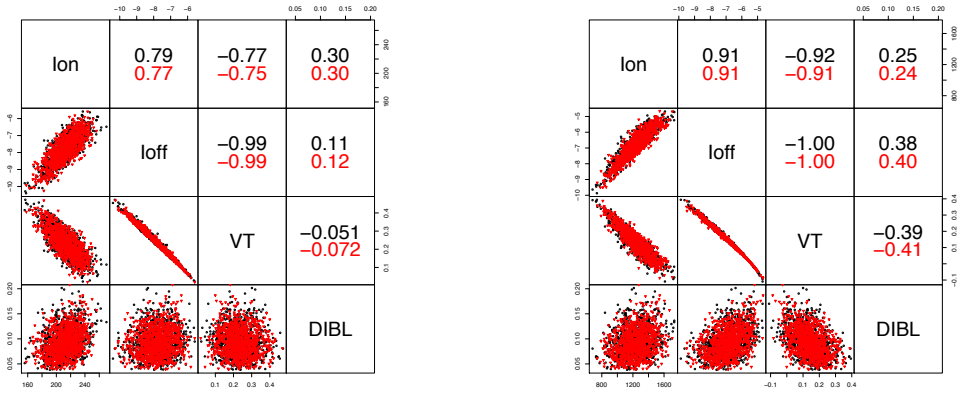


Fig.5.29 Correlations of figures of merit between physical simulation and compact model generator for NMOS devices at ageing time t=3 month. The left figure is when V_{DS}=0.05V, while the right figure is when V_{DS}=1V.

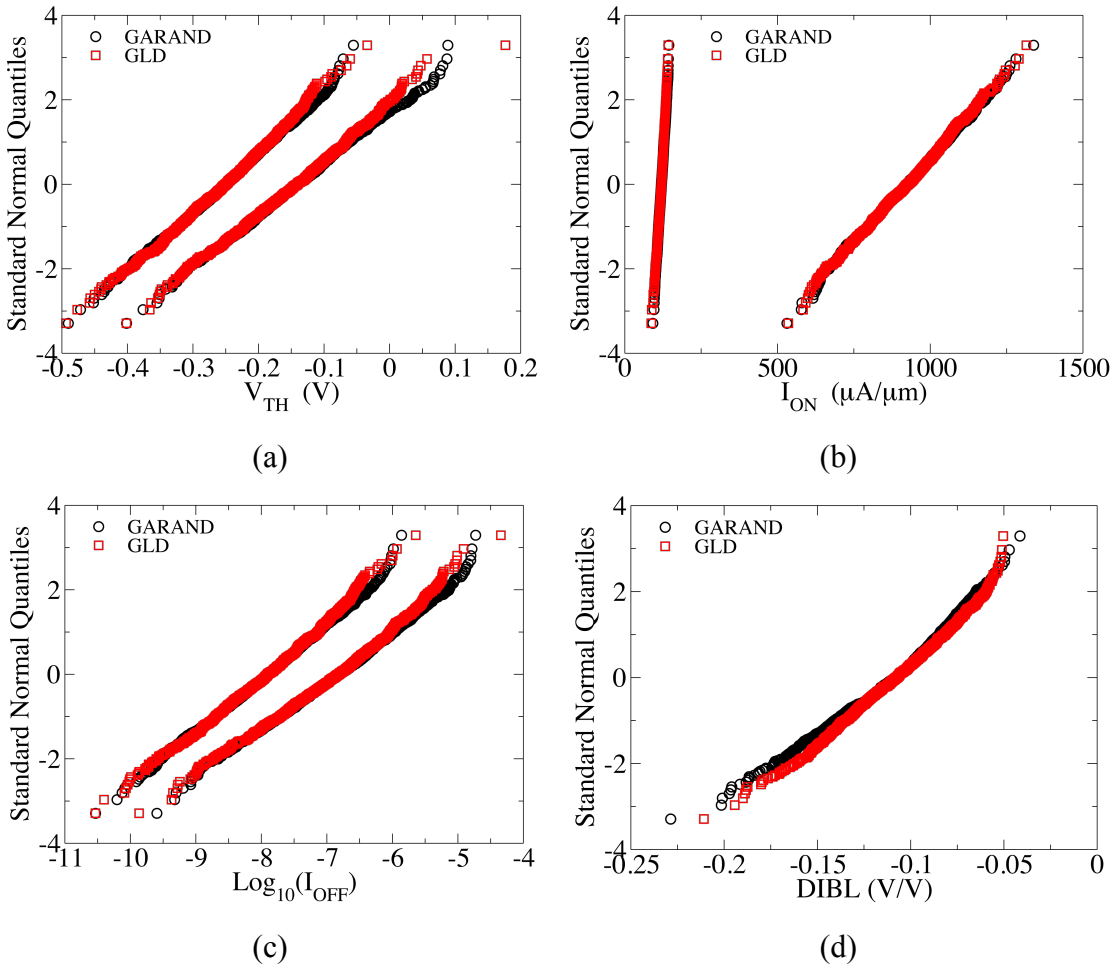


Fig.5.30 Comparisons of figures of merit (a)V_{TH} (b)I_{ON} (c)I_{OFF} (d)DIBL of PMOS between physical simulations and compact model generator at ageing time t=3 month.

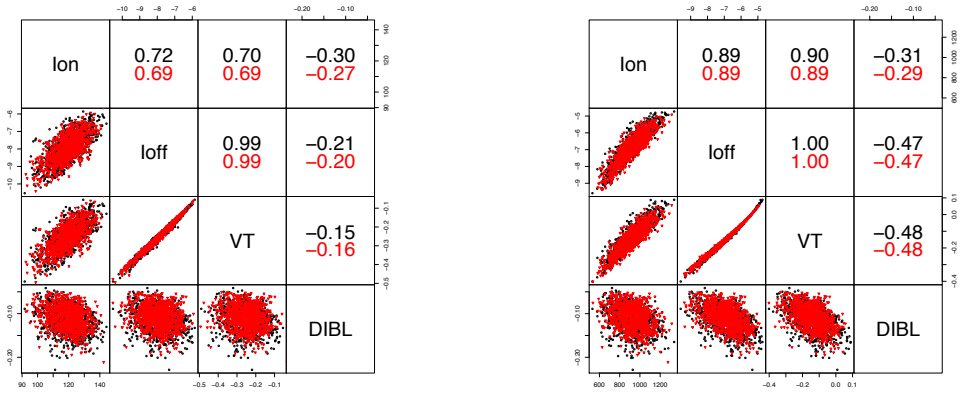


Fig.5.31 Correlations of figures of merit between physical simulation and compact model generator for PMOS devices at ageing time t=3 month. The left figure is when V_{DS}=0.05V, while the right figure is when V_{DS}=1V.

We expect the distribution of figures of merit follows the trend of TCAD simulation. Fig.5.32 shows V_{TH} distribution of 100,000 devices from GLD and Gaussian V_T against the TCAD simulation result. From this picture, we can see that the distribution of V_{TH} from GLD captures the body and follows the trend. However, Gaussian V_T generated devices start to deviate from the GLD devices into both tails of the distribution. If devices are generated for a higher sigma investigation, the deviation of Gaussian V_T approach will be more obvious. These 'tail devices' are critical when running high-sigma analysis and when attempting to estimate yield, and its important that the generation methodology is managing to reproduce these devices.

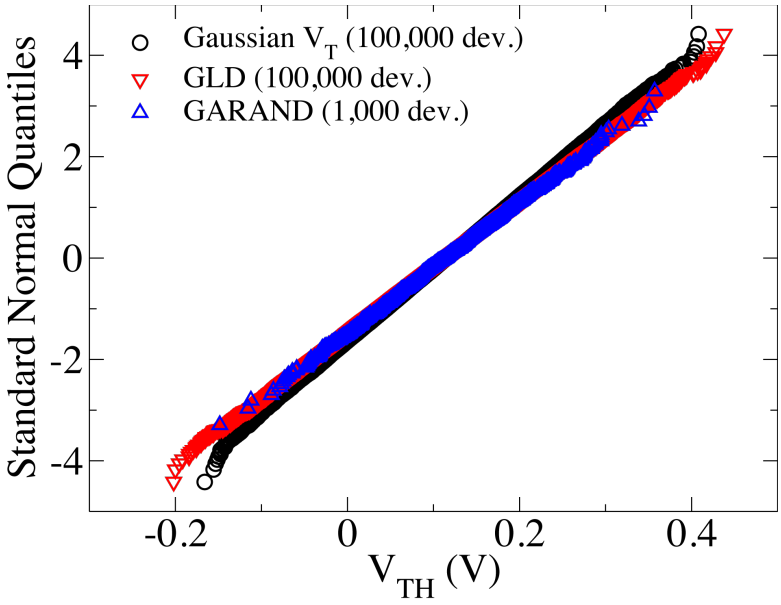


Fig.5.32 V_{TH} comparison of fresh devices when V_{DS}=1V.

5.8 Summary

This chapter mainly addresses, illustrates and verifies compact model generation methodologies, by which a sufficiently large number of accurate compact models can be generated at any arbitrary ageing time. The GLD method enables compact model parameter generation "on the fly". The trap density interpolation method assists GLD to generate compact models at arbitrary trap density. The ageing model translates trap density to ageing/stress time, which greatly benefits the circuit designer. The methodologies above are successfully embedded into the SPICE simulator RandomSpice.

The accuracy of the compact model generation methodologies are validated by comparing the new generated compact models against the physical simulation. In this chapter, 100,000 compact models generated by GLD are compared with the Gaussian V_T method, showing that GLD method follows the trend of the physical simulation more accurately. GLD is better for compact model generation and especially at high sigma investigation and provides more accurate information of the tailed devices.

The compact model generation methodology enables the investigation of the influence of statistical variability and BTI-induced ageing at circuit level. In Chapter 6, the generated compact models at different ageing times will be applied on the 6T SRAM to investigate the influence of different stages of ageing on SRAM stability and write performance.

Chapter 6

Statistical SRAM simulation

6.1 Introduction

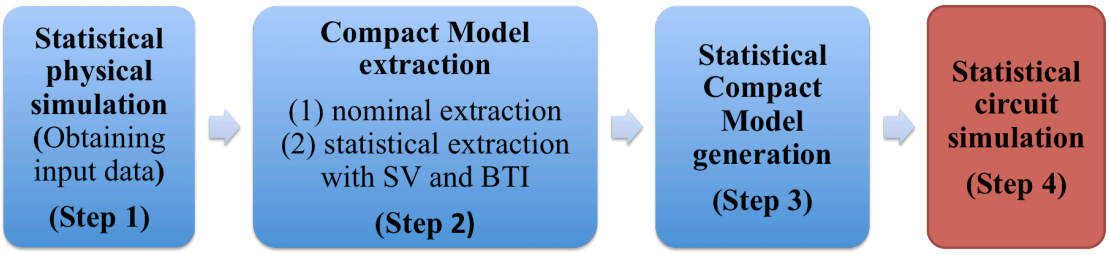


Fig.6.1 Step 4 of the simulation flow.

In Chapter 5, the methodology is demonstrated and verified enabling the generation of a sufficiently large number of compact models with statistical variability at any arbitrary ageing time, and is successfully propagated into SPICE simulator RandomSpice. This chapter focuses on the last step of the simulation flow, statistical circuit simulations (shown in Fig.6.1 in red block). In this chapter, the generated compact models are applied to a demonstrator circuit – a standard 6T SRAM cell. The influence of ageing on SRAM stability and dynamic write performance is investigated. As variations induced by statistical variability and BTI-induced ageing are stochastic, large-scale statistical circuit simulations need to be performed for the investigation. High sigma investigation for SRAM is available using the compact model generating approach. This chapter aims to propagate the variability and reliability data generated by the TCAD simulations to

evaluate the impact on SRAM circuit performance in respect to both variability and ageing.

In section 6.2, the motivation for using SRAM as the test-vehicle to investigate statistical variability and BTI-induced ageing is introduced, followed by the introduction of SRAM metrics, including Static Noise Margin (SNM) and dynamic Write Margin (WM). Section 6.3 presents the sensitivity SNM and WM to each individual transistor's variation. Section 6.4 demonstrates the dependence of SNM and WM on ageing effects and analyses the underlying reasons. In section 6.5, the response surface of SNM and WM are shown, which gives a direct view of the change of SNM and WM. Finally, the research in this chapter is summarized in section 6.6.

6.2 SRAM As the Vehicle

SRAM is frequently used as a benchmark of a particular technology node [105]. An SRAM array is always adopted for microprocessor caches as an integral component due to its read/write speed. Due to the limited integration area, the corresponding SRAM cell usually consists of devices with minimum dimensions for a technology node. For example, SRAM under Intel 65 nm technology containing >0.5 billion transistors only takes up 0.57μm^2 [106]. Therefore, compared to other circuits, SRAM is the circuit that is most vulnerable to statistical variability and BTI-induced ageing, as it is composed of minimum geometry devices, and it relies on matched transistor performance [107]. SRAM always has very strict design constraints on chip area as well as on power density. Within an SRAM block, a single cell failure can be catastrophic by damaging a complete array if redundancy is not used. Due to the trade off between periphery control circuit area and complexity, and the redundant array area, redundancy is used for the entire row or column in which the failed cell is located rather than replacing the single cell. SRAM yield is particularly challenging. For an SRAM row or column with N cells, if the failure probability for each cell is p and the failure is independent, the entire row or column failure probability will be

$$P = 1 - (1 - p)^N \tag{6.1}$$

For example, if each cell holds the failure of 0.001, then for the entire row containing 100 cells, the failure probability is 0.095. Therefore, it is critical to investigate each

single SRAM cell performance under the influence of statistical variability and ageing. In this study, SRAM was chosen as the test vehicle. With the developed approach in Chapter 5, it is possible to investigate the impact of BTI-induced ageing on 6-T SRAM performance.[102]

Fig.6.2 shows the diagram of 6T SRAM, which consists of 6 transistors, including two pull-up PMOS transistors on the left (PUL) and right (PUR) sides, two pull-down NMOS transistors on the left (PDL) and right (PDR) sides, two pass gate NMOS transistors on the left (PGL) and right (PGR) sides. The following channel widths are used for the three type of transistors in the cell: pull-up (PU) transistors 50 nm, pass gate (PG) transistors 75 nm and pull-down (PD) transistors 150 nm. The peripheral circuit is also employed in this study to generate clock signals for read and write performance, ensuring the accuracy of SRAM access operation. Overall, we expect the periphery circuit to slow down as a function of time. This will have a different impact on different SRAM figures of merit. For example, we could expect the word line pulse width to increase, thus aiding write and read operations and potentially helping to cancel out some of the detrimental effects of BTI on the bitcell, however it will adversely affect the access stability. The impact of BTI on the periphery circuit is not included in this work as the aim of this chapter is to focus on the SRAM cell.

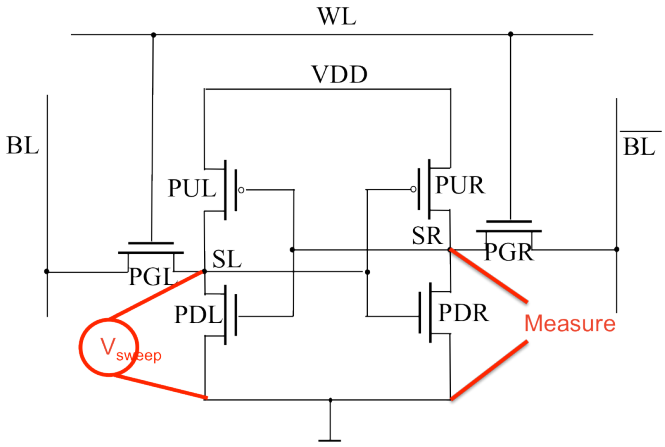


Fig.6.2 6T SRAM schematic view.

The impact of statistical variability and BTI-induced ageing on SNM is evaluated in this research. SNM is measured by connecting a voltage source to one of the internal cell nodes. The applied bias from the voltage source is then swept from 0V to Vdd and the

voltage at the opposite internal node is measured, shown in Fig.6.2. The BL and $\overline{BL}$ are all at '1' and PGs are opened during the measurement. This procedure is then repeated for the opposite internal node. The largest square that can be fitted within each of the two loops of the butterfly curve is calculated and the SNM of the cell is then defined as being the smaller of these diagonals, as shown in Fig.6.3. When this diagonal becomes too small, it becomes difficult or impossible for the cell to retain its state, and the value stored in the cell is lost [108].

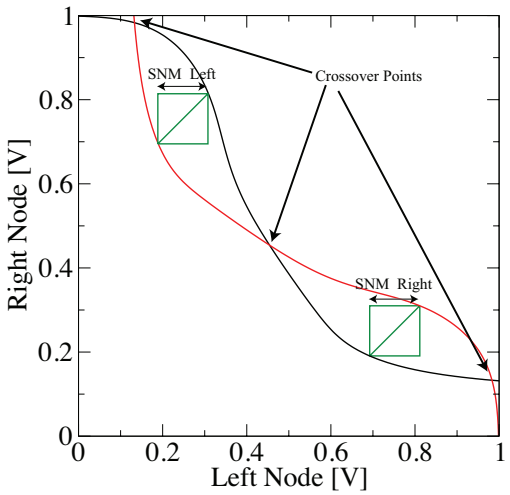


Fig.6.3 SNM definition for a ballanced SRAM cell (SNM left = SNM right) [102].

We also evaluate SRAM dynamic performance by measuring WM; defined as the time between the rising node reaching 70% of Vdd, and the word line (WL) falling to 50% of Vdd in one write circle (shown in Fig.6.4)). The write WL high pulse time in this study is 165ps at the chosen simulation Process-Voltage-Temperature (PVT) corner, which is a common industrial standard. A negative WM or failed measurement indicates that the write operation has failed, while a positive WM is the time remaining after a successful writing operation. The larger the margin is, the faster the cell can be written to. The smaller the margin is, the more difficult it is to write into the cell.

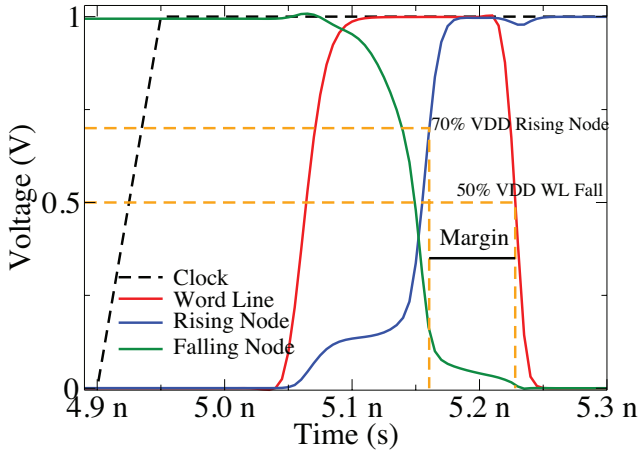


Fig.6.4 Dynamic Write Margin definition [109].

6.3 Sensitivity Analysis

In this section, a sensitivity analysis is carried out and the results are discussed. The sensitivity test is performed by firstly injecting statistical variability into a single transistor in the cell while the remaining transistors are kept uniform, and then measuring the circuit metric response. This procedure is repeated for each transistor in the circuit. For each transistor under statistical variability, 1,000 SRAM cells are simulated for SNM and WM respectively, so that we can estimate the impact of one single device variability on the chosen cell metric. For example, we take PGL as the device presenting statistical variability. Standard deviation is calculated from the 1,000 simulation results, which shows SNM and WM deviations induced by the variations from PGL. If the value of the standard deviation is high, that particular metric is very sensitive to the PGL variations.

Table 6.1 shows the sensitivity test of SNM results. As SNM is calculated based on symmetrical simulations (both the left and right nodes are swept), we can see that over the full statistical ensemble, the left and right devices in the same position have an equivalent sensitivity. As a result, we can analyse the transistors in pairs – PGL and PGR, PUL and PUR, PDL and PDR. From this table, we can also see that PDs are the most sensitive transistors to SNM, since they hold the highest standard deviation of 0.013V, followed by the PG transistors (0.006V) and PU transistors (0.003V).

Table 6.1 Sensitivity test SNM results.

Transistor Under Test	Standard Deviation of SNM (V)
PGL	0.0059
PGR	0.0059
PUL	0.0032
PUR	0.0030
PDL	0.0135
PDR	0.0131

WM is measured when writing '1' to the left node of the cell, which is initialised to '0'. Transistor sensitivity results for WM are shown in Table 6.2. In the result, PGR takes the highest value of standard deviation (5.79ps). This means that PGR's variation has the strongest influence on WM, followed by PUR, PGL, PDL, PUL, PDR. The sensitivity test results will be discussed with the combination of the results in the following sections.

Table 6.2 Sensitivity test WM results.

Transistor Under Test	Standard Deviation of WM (ps)
PGL	2.33
PGR	5.79
PUL	1.23
PUR	2.99
PDL	2.11
PDR	1.79

6.4 SRAM Simulations

In order to investigate the effect of statistical variability and BTI-induced ageing, three ageing scenarios are investigated. In the normal situation, PU/PD devices experience the frequent stress and relaxation process, but PGs have short word line 'ON' time. For the first scenario, we assume PU and PD transistors in the cell age uniformly. This aims to represent a situation where data is frequently changed in cell. However, the rate of transistor ageing depends on the time for which each transistor in the cell is subject to electrical or thermal stress. Therefore, transistors in the same cell may age at different

speeds due to different operating patterns and the mismatch between the inverters can increase – this is investigated in the second scenario. In this scenario, we focus on the change in WM when the cell remains in one state (storing '0' or '1') for a long period of time, resulting in ageing mismatch between the two cross-coupled inverters. In the third scenario, we investigate the effects of ageing on the PG transistors, which play an important role in SRAM performance, by determining the timing of both write and read operations. In each of the scenarios above, 10,000 SRAM cells are simulated in order to investigate the influence of ageing and statistical variability on cell stability and write performance. The ensemble size of 10,000 gives low standard errors on mean (1.01%) and standard deviation (0.71%), ensuring the circuit performance prediction accuracy. Fresh transistors, transistors at low, medium and high ageing levels that translates to trap densities are $1x10^{11}cm^{-2}$, $5x10^{11}cm^{-2}$, $1x10^{12}cm^{-2}$ are used in these scenarios. Using these ageing levels can clearly tell SRAM performing trend when the cells age gradually. Finally, the response surfaces of SNM and WM are shown when different parts of transistors age differently.

6.4.1 Transistors Age Uniformly

If data change frequently, we can assume that degradation in all PU and PD transistors is relatively balanced. This is the most optimistic measure of the impact of BTI on the bitcell because it does not lead to any mismatch in the cell. This could be true in a subsystem like a level 1 (L1) cache where, due to the demand for data from the CPU, data in the cache are often refreshed. Thus, stored data can be changed with high frequency in a L1 cache and the active time of the opposing PU and PD transistors can be nearly equivalent, resulting in the same ageing rate of PU and PD transistors. We will focus on this situation in this section while the simulation scenario is arranged such that the PU and PD transistors, on both sides of the cell, age uniformly from fresh to the highest level of ageing (N_t=1×10^{12} cm^{-2}). The relatively slow ageing of PG transistors is ignored in this scenario, in order to fully investigate the influence that PU and PD transistors' ageing has on SNM and WM. Fig.6.5 shows the corresponding SRAM diagram with the red circles indicating the aged transistors.

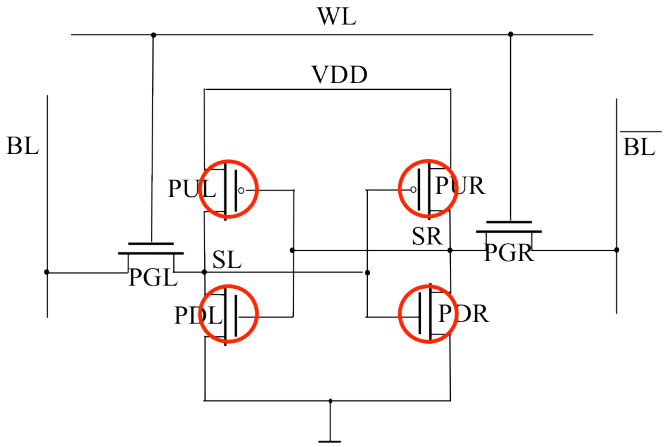


Fig.6.5 SRAM diagram when two inverters age uniformly.

The SNM results are shown in Fig.6.6, Fig.6.7, and Fig.6.8. From QQ plot in Fig.6.6, we can directly see that the 10,000 simulated SRAM cells' SNM are widely distributed, which is also reflected in Fig.6.7 on the standard deviation value. For the cells with only fresh transistors, the mean of SNM is 0.11 and the standard deviation is approximately 0.017, which shows an average of 15% variation for each cell deviating from the mean value. As the trap density of PU and PD transistors increases, the standard deviation increases as well, showing that the more variations on PU and PD transistors, the more variations on the cell's stability. Additionally, the increase of PU and PD's degradation level makes average SNM slightly increase respectively by the steps of 0.24%, 1.81%, 2.96%, respect to the fresh cell. The reason behind it can be obtained by analysing the comparison of butterfly curves from the fresh cell and the cell with PU and PD at the highest ageing level respectively in Fig.6.8. Focusing on sweeping the left node (SL) and measuring the right node (SR), it is clear to see that the curve move towards right after degradation compared with the curve presenting the fresh cell, showing it becomes difficult for the right node to approach to '0' after ageing. The degradation on PDR transistor reduces its strength to pull down the right node to '0', showing that the curve should move right. However, the PUR transistor's ability to pull up the right node to '1' is also reduced after degradation, showing that the curve should move left. The combination of these two cases is that the curve moves right (Fig.6.8), showing that it becomes harder to change the state, also showing that the degradation on PDR is more dominant than PUR. This is also agreed with the sensitivity test for SNM in section 3.1, that PD transistors hold much higher value (0.0131) than PU transistors (0.0030). The conclusion is the same for sweeping the right node (SR) and measuring the left node

(SL). Therefore, in Fig.6.8, the two red curves, representing the cell performance after degradation, move right, resulting in the new balance of the cell and the slightly changed SNM.

When the left node (SL) sweeps to '1', the end of the red curve is slightly higher than the black curve, indicating after sweep, the voltage at the right node (SR) is higher than the voltage of the un-aged condition. Transistors hold higher resistance after degradation. The voltage at the end of the curve depends on the resistance ratio between PGR and PDR, since PGR and PDR connect the $\overline{BL}$ ('1') to the ground. The resistance of PDR is higher after degradation, however the resistance of PGR does not change. Therefore, the voltage on PDR is higher than before degradation, and the end of the curve moves up slightly.

Here it should be noted that the curves sweeping from the opposite sides are not totally symmetric since the transistors are randomly generated at the particular ageing level, and do not hold the same performance. It should also be noted that the non-Gaussian distribution of SNM from Fig.6.6 is due to the nature on calculating that the value of SNM is not choosing from the certain side of the butterfly curve, but the smallest diagonal of the two largest squares fitted in each side of the butterfly curve. This also explains all the non-Gaussian distributions of SNM in the following sections.

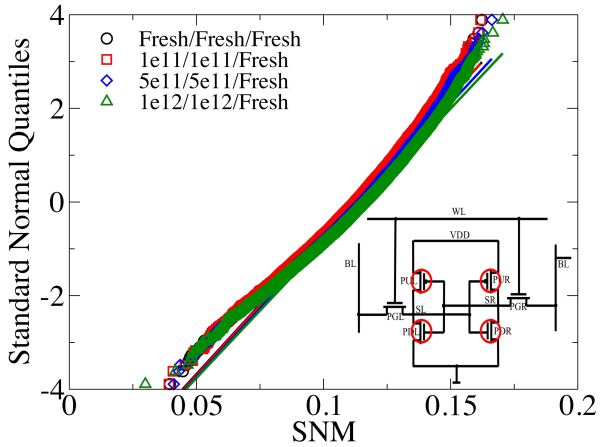


Fig.6.6 QQ plot of SNM when when PU and PD transistors age uniformly. The legend is the ageing level of (PUR and PDL)/(PUL and PDR)/PG transistors.

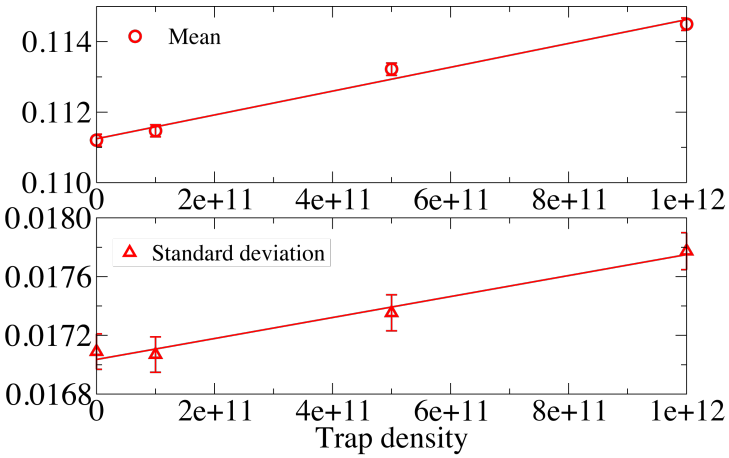


Fig.6.7 Evolution of the average SNM and its standard deviation when PU and PD transistors age uniformly.

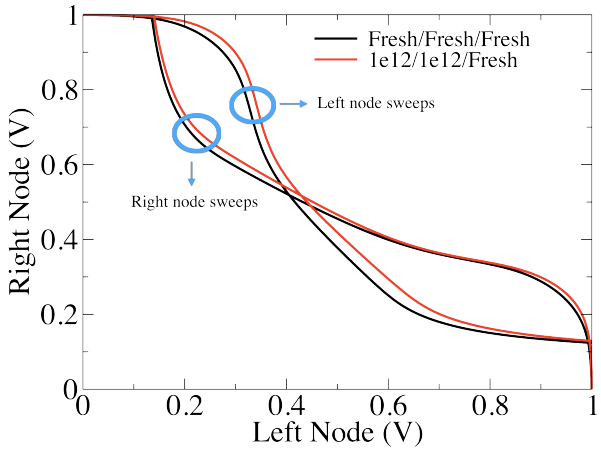


Fig.6.8 Butterfly curves of the fresh cell and the cell with PU and PD transistors at the highest ageing level respectively. The legend is the ageing level of (PUR and PDL)/(PUL and PDR)/PG transistors.

The WM results are shown in Fig.6.9 and Fig.6.10. Fig.6.9 shows the distribution of WM and Fig.6.10 shows the evolution of the average WM and its standard deviation with ageing. Compared with the fresh cells, the mean of the WM of cells at low, medium and high ageing levels increases 0.46%, 2.23% and 4.43%, while standard deviation decreases 0.53%, 2.47% and 4.55% respectively. WM increases as trap density increases, showing that writing becomes faster. As WM increases, write time reduces, as a result, the standard deviation of WM reduces. In the following sections, the same reason applies, that WM increases, but standard deviation decreases. The SRAM write operation is performed by writing a '0' at the side where a '1' is stored, meaning on that side the PG transistor must overcome the PU transistor. In this case, the PGR transistor needs to overcome the PUR transistor in order to complete the write cycle. This becomes easier

due to the reduced pull-up strength of the PUR and results in an increase in WM. Thus, regardless of writing a '0' or '1' to the SRAM cell, when PU and PD transistors are uniformly aged, they have a positive impact on write performance and the WM increases.

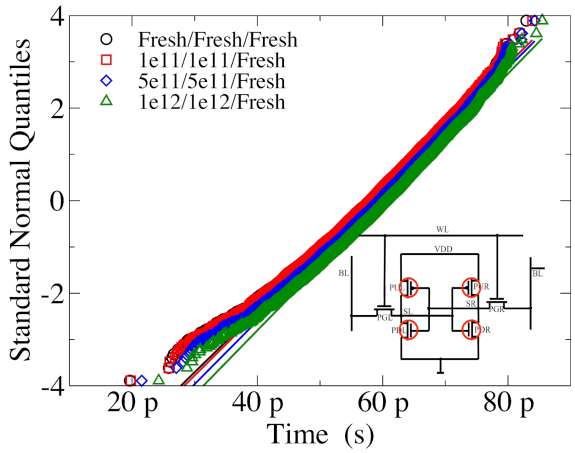


Fig.6.9 QQ plot of WM when PU and PD transistors age uniformly. The legend is the ageing level of (PUR and PDL)/(PUL and PDR)/PG transistors [109].

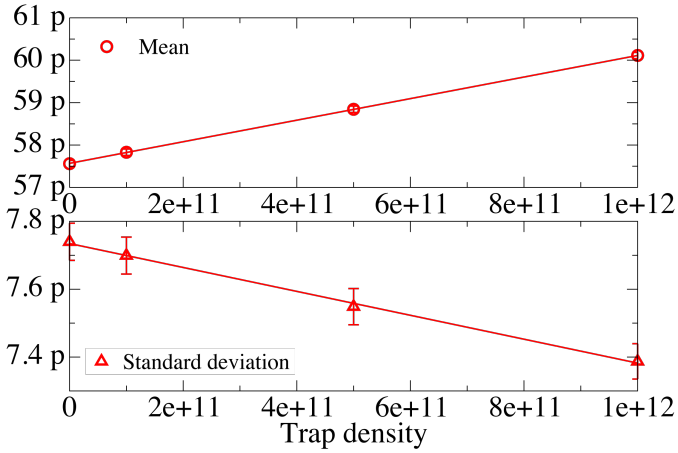


Fig.6.10 Evolution of the average WM and its standard deviation when PU and PD transistors age uniformly [109].

6.4.2 Mismatch Between the Two Cross-coupled Inverters.

In the second scenario, where the cell remains in one state for a long period of time or the same data is repeatedly written, a transistor of each inverter in the cell is constantly on (called 'ON' transistors) while the other is off. The corresponding two 'ON' transistors

age faster than others, by being in the BTI stress conditions and therefore their threshold voltage increases. This results in an increase of the mismatch between the two inverters in the SRAM. For example, if in the 6T cell in Fig.6.11(a) the SL side constantly holds '0', the PDL and PUR transistors are always at the 'ON' state and age faster than the other transistors in the cell. In this section, we assume the extreme situation that 'ON' transistors age from fresh to the highest level of degradation while other transistors stay fresh. In this way, the effect of the mismatch between the two cross-coupled transistors on SRAM performance is maximized and we can understand the worst-case scenario.

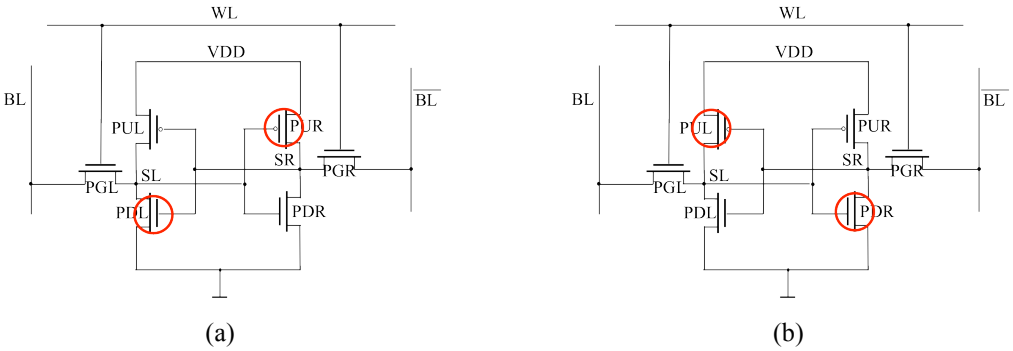


Fig.6.11 SRAM diagram with mismatch. In (a), PDL and PUR are aged. In (b), PUL and PDR are aged.

Beginning with SNM, the corresponding results according to Fig.6.11 (a) are shown in Fig.6.12, Fig.6.13, and Fig.6.14. From these pictures, we can see that the mean value of SNM decreases with the increase of ageing. The standard deviation increases as ageing increases, showing that the increased ageing on PDL and PUR makes the SNM more widely distributed. The average decreasing percentages of mean of SNM with respect to the fresh cell are 0%, 0.78%, 4.74% at ageing level of low, medium and high respectively. In order to find the reason behind this, we focus on the butterfly curve comparison between the fresh cell and the cell with PUR and PDL aged at the highest ageing level, which is shown in Fig.6.14. First, we focus on when the left node (SL) sweeps from '0' to '1'. The upper part of the curve moves towards left after degradation, meaning that the right node (SR) is faster to fall to '0'. This is due to the degradation of PUR, which reduces the pull-up strength. However, the lower part of the curve only slightly moves towards left, and near the end of the curve, it does not change after degradation. This is because at the lower part of the curve, PDR transistor is opened.

Though ageing of PUR helps PDR to pull down the right node (SR) to '0', compared with PDR's pull-down ability, the influence of ageing on PUR can nearly be neglected.

Focusing on sweeping the right node (SR) from '0' to '1', in this situation, PDL is aged. The whole curve after degradation drops slower to '0' due to the degradation of PDL, that its ability to pull down SL to '0' is reduced. However, we can see that the right part of the curve drops much slower to '0' compared with the black curve presenting un-aged condition, indicating that the influence of ageing on PDL is not negligible when PUL is opened. Comparing with the negligible influence of degradation on PUR stated above, it shows again that when sweeping one side of the internal node, between the PU and PD transistors on the measuring side, PD transistor is more dominant in changing the state of the measuring side.

From the analysis above and also shown in Fig.6.14, one side of the butterfly curve becomes smaller while the other becomes larger. However, the SNM is determined by the smallest diagonal of the two largest squares fitted in each side of the butterfly curve. Thus, the SNM decreases in this situation.

When the PUL and PDR are aged and others remain fresh (shown in Fig.6.11 (b)), the corresponding results are shown in Fig.6.15, Fig.6.16, and Fig.6.17, indicating the decrease of SNM as PUL and PDR's ageing level increases. The decrease percentages compared with fresh cell are 0%, 0.84%, 4.84% at ageing level of low, medium and high respectively. The similar decreasing speed between situations of Fig.6.11 (a) and (b) supports the conclusion drawn in SNM sensitivity analysis, that the left and right devices in the same position have an equivalent sensitivity to SNM. Fig.6.17 shows the corresponding butterfly curve in the situation of Fig.6.11(b). However, comparing with the butterfly curve in Fig.6.14, it can be seen that the butterfly curve shown in Fig.6.17 is the reverse of the curve in Fig.6.14. This is because only the sweep and measure points are swapped between the two situations. Hence, they hold the same result that SNM decreases as the mismatch increases.

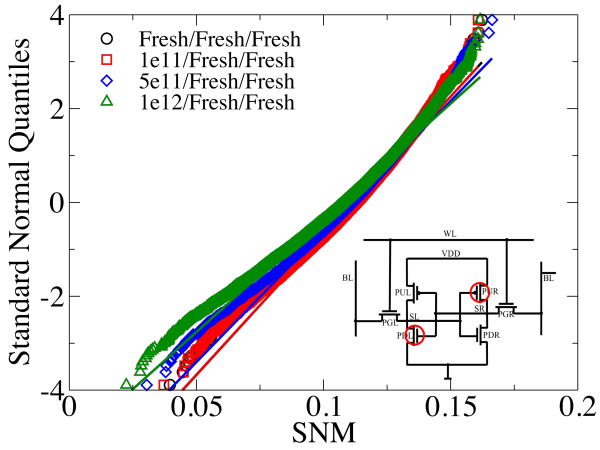


Fig.6.12 QQ plot of SNM when PDL and PUR transistors are degraded. The legend is the ageing levels of (PUR and PDL)/(PUL and PDR)/PG transistors.

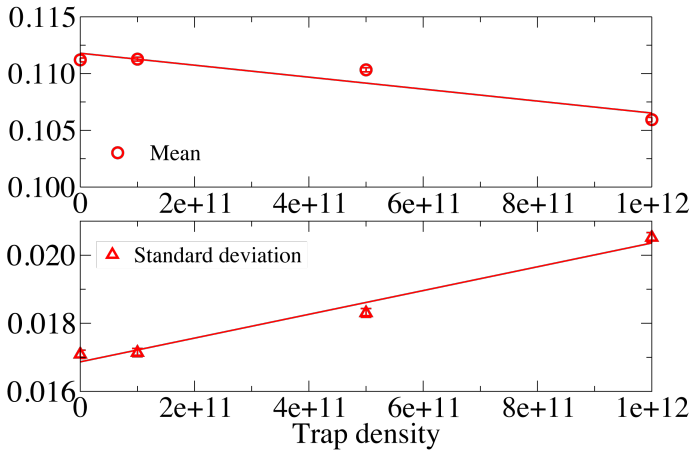


Fig.6.13 Evolution of the average SNM and its standard deviation when PDL and PUR transistors are degraded.

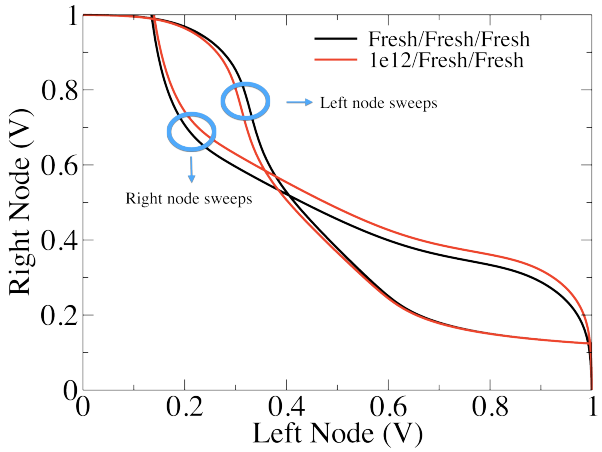


Fig.6.14 Butterfly curves of the fresh cell and the cell with PDL and PUR transistors at the highest

ageing level respectively. The legend is the ageing levels of (PUR and PDL)/(PUL and PDR)/PG transistors.

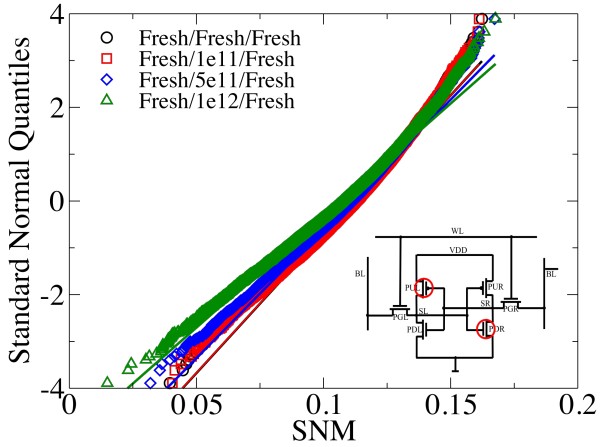


Fig.6.15 QQ plot of SNM when PUL and PDR transistors are degraded. The legend is the ageing levels of (PUR and PDL)/(PUL and PDR)/PG transistors.

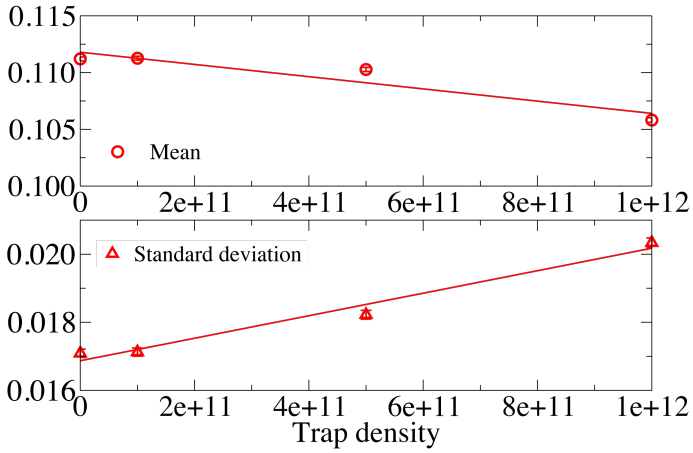


Fig.6.16 Evolution of the average SNM and its standard deviation when PUL and PDR transistors are degraded.

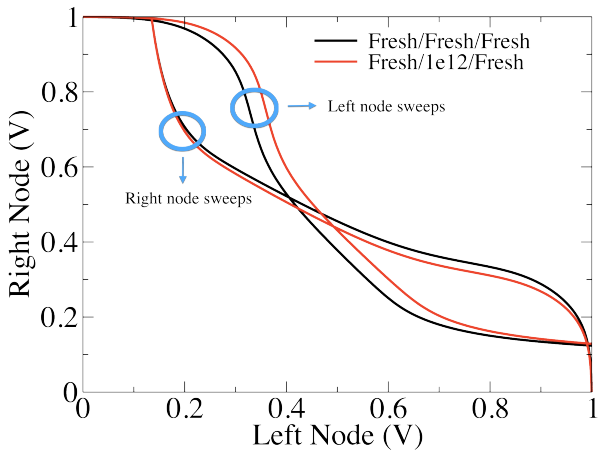


Fig.6.17 Butterfly curves of the fresh cell and the cell with PUL and PDR transistors at the highest ageing level respectively. The legend is the ageing levels of (PUR and PDL)/(PUL and PDR)/PG transistors.

For WM, we consider the case where the PDL and PUR transistors are the 'ON' transistors as shown in Fig.6.11 (a) and a '1' is written to the SL side firstly. Fig.6.18 is the corresponding evolution of WM distribution. Fig.6.19 presents the average WM and its standard deviations with ageing.

In Fig.6.18 and Fig.6.19, it is clear that WM increases as the mismatch increases. The average increase is 1.10%, 5.23% and 10.02% compared to the fresh cell. It is obvious that ageing of the PUR transistor helps writing '0' to the right side of the cell, which was previously discussed in section 6.3.1. Ageing of PDL also benefits WM as the increased V_{TH} in PDL makes this transistor easier to turn off, which is beneficial to write '0' to the SR side.

However, comparing with Fig.6.9, it is clear that the increase in WM in Fig.6.18 is more pronounced, indicating that ageing on PUL and PDR transistors makes writing '1' to the SL side more difficult. This is due to the fact that ageing in PUL and PDR transistors reduces their ability for the PUL to pull up SL at '1' and for the PDR to pull down SR to '0'. We verified the explanation by swapping 'ON' transistors from PDL and PUR to PUL and PDR (shown in Fig.6.11 (b)), the results of which are shown in Fig.6.20. Fig.6.20 shows the WM distribution while Fig.6.21 shows the corresponding evolution of the average WM and its standard deviation as ageing increases. WM decreases when the

ageing level of PUL and PDR increases, with a decrease of 0.63%, 3.04% and 5.75%, compared with fresh cells.

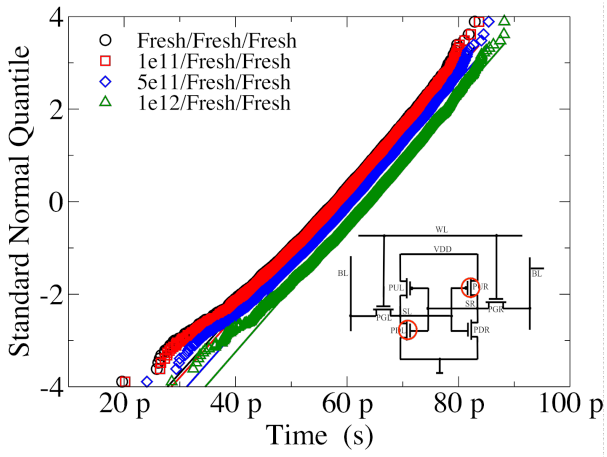


Fig.6.18 QQ plot of WM when PDL and PUR transistors are degraded. The legend is the ageing levels of (PUR and PDL)/(PUL and PDR)/PG transistors [109].

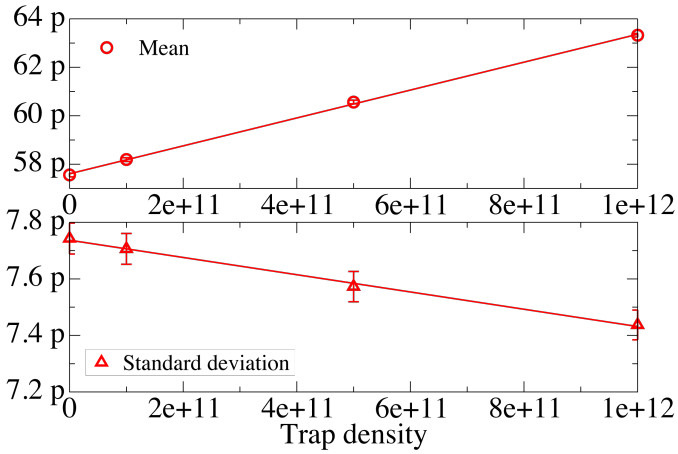


Fig.6.19 Evolution of the average WM and its standard deviation when PDL and PUR transistors are degraded [109].

Compared with the increase in WM observed in Fig.6.18, the decrease shown in Fig.6.20 is much smaller when the ageing of PUL and PDR increases. This matches the results shown in Fig.6.9, which combined these two situations (all PU and PD transistors are degraded) that WM increases but significantly less. The conclusion can be easily drawn, stressing that the increase influenced by PDL and PUR's ageing is greater than the decrease brought by the PUL and PDR' ageing. Writing is performed by changing the side where a '1' is stored to '0'. In this case, the PGR transistor needs to overcome the PUR transistor to finish writing. This highlights the importance of PUR in the writing

performance. This also matches the sensitivity test for WM that, PUR holds the highest sensitivity value in the four transistors in the middle and thus has the strongest effect on WM (refer to Table 6.2). Therefore, PUR (the PU transistor holding '1' before the write) plays the most important role between the two inverters for the write performance.

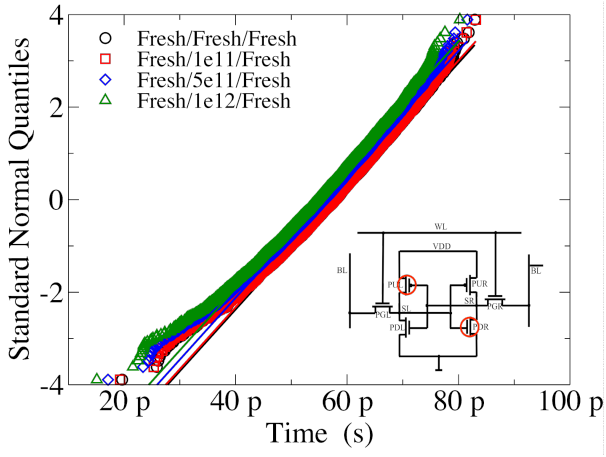


Fig.6.20 QQ plot of WM when PUL and PDR transistors are degraded. The legend is the ageing levels of (PUR and PDL)/(PUL and PDR)/PG transistors [109].

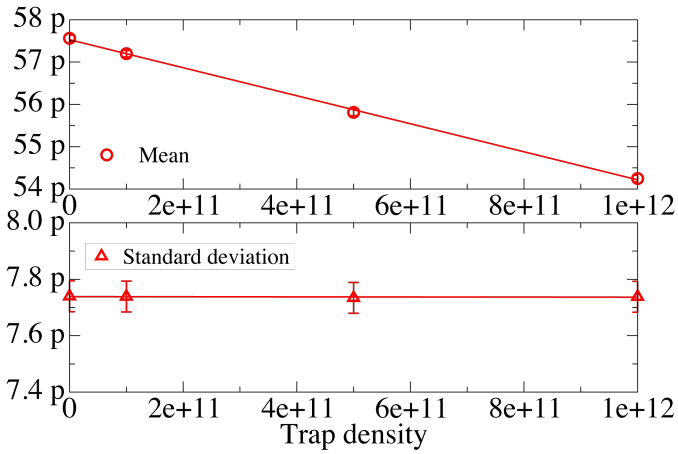


Fig.6.21 Evolution of the average WM and its standard deviation when PUL and PDR transistors are degraded [109].

Meanwhile, mismatch between two inverters can be either positive or negative for write performance, depending on whether the aged transistors are turned on or off after writing. If they are turned on, the WM is decreased (Fig.6.20), while if they are turned off, the WM is increased (Fig.6.18). Indeed, the aged transistors are easier to turn off and more difficult to turn on compared with the fresh transistors. Therefore, for an SRAM cell, as the mismatch increases, writing to one state becomes faster while writing to the

other state becomes slower. This enlarges the time difference between writing '0' and '1'. Mismatch improves one side of the write performance and degrades the opposite side.

6.4.3 Impact of Ageing on PG Transistors

PG transistors are subject to stress regardless of whether the cell is in read from or written to. However, word line pulses are relatively short, any single word line is only active for a very small percentage of the overall memory operation. When the word line is '0' but bit line is pre-charged, the PG transistor on the internal '1' side has potential difference between its gate & source and gate & drain. So PG transistor on the '1' side will have two-sided gate leakage. Similarly, the PG transistor on the internal '0' side will have one-side gate leakage [110]. In this section, ageing impact on PG transistors is investigated, shown in Fig.6.22. Since PG transistors age much slower than PU and PD transistors, the simulation scenario is arranged that PU and PD transistors' ageing level is fixed at $1x10^{12}cm^{-2}$, while PG transistors' ageing level increases from 0 to the highest. This represents an extreme degradation scenario for the PGs, but is beneficial to understand the limiting case of PG ageing.

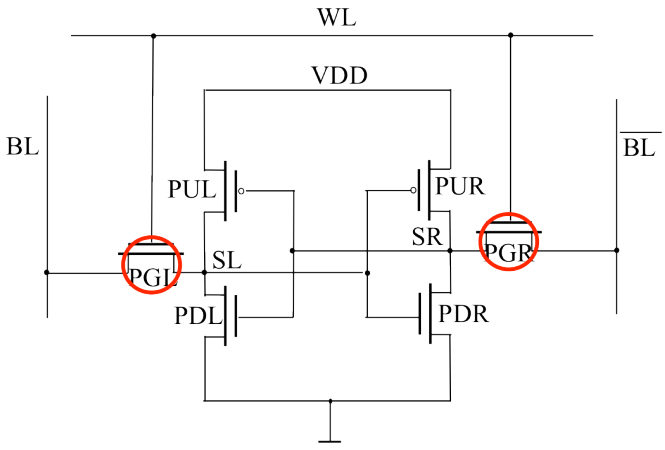


Fig.6.22 SRAM diagram for the investigation of ageing impact on PG transistors.

Fig.6.23 shows the distributions of SNM when PG transistors age from 0 to the highest level. Fig.6.24 is the corresponding evolution of the average SNM and its standard deviation with ageing. Fig.6.25 shows the butterfly curves of the cell with fresh PGs and with PGs at the highest ageing level. When PG transistors age, SNM increases at the

speed of 1.03%, 5.30%, 10.43% compared with the fresh cell. This matches the results in Fig.6.25, that after PGs are aged, any side of the butterfly curve becomes large, indicating that the cell becomes more stable. During the SNM measurement, BL and $\overline{BL}$ are both at '1' and PG transistors are opened. The aged PG on the measurement side reduces the influence of the BL (or $\overline{BL}$), to the internal measuring node that will fall to '0' during the sweep. For example, when sweeping the left node (SL) from '0' to '1', as PG transistors' ageing level increases, the strength of PGR transistor to pull up the right node (SR) is greatly reduced, and it becomes easier for PDR to pull down to '0' when the PDR is turned on. Correspondingly, the upper part of the curve does not change and the lower part of the curve moves left. For sweeping the right node (SR), the result is the same. Therefore, both sides of the butterfly curve become larger and the SNM increases.

Meanwhile, at the end of the sweep, the red curve is lower than the black curve (shown in Fig.6.25). The voltage at the end of the curve depends on the resistance ratio between PGR (connecting to '1') and PDR (connecting to '0'). The resistance of PGR is higher after degradation, however the resistance of PDR does not change. Therefore, the voltage on PDR is lower than before degradation.

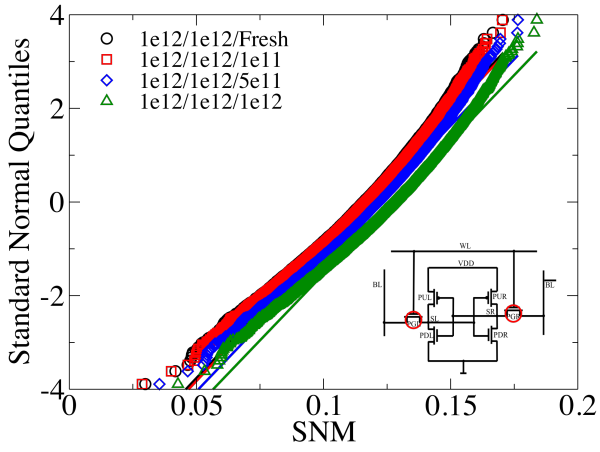


Fig.6.23 QQ plot of SNM when PG transistors are degraded. The legend is the ageing levels of (PUR and PDL)/(PUL and PDR)/PG transistors.

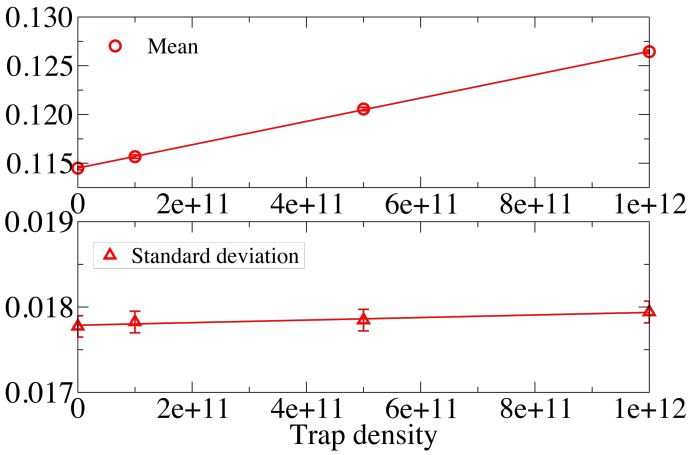


Fig.6.24 Evolution of the average SNM and its standard deviation when PG transistors are degraded.

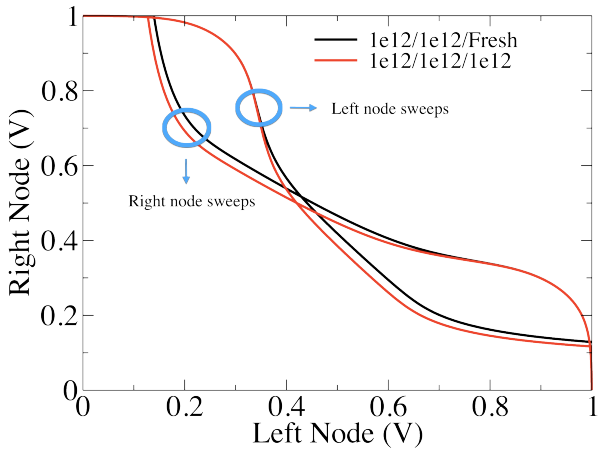


Fig.6.25 The butterfly curves of the cell with fresh PGs and with PGs at the highest ageing level. The legend is the ageing levels of (PUR and PDL)/(PUL and PDR)/PG transistors.

Fig.6.26 shows the corresponding evolution of the WM distribution. Fig.6.27 shows the corresponding evolution of the average WM and its standard deviation with ageing. A significant decrease occurs (average WM decreases from 60.1ps to 51.7ps), and WM even decreases 13.99% for the highest level of degradation compared to the fresh cells. This is a non-negligible drop, showing the impact of ageing in PG transistors on write performance. PG transistors directly determine the access time and if aged, the write time and read time will be increased due to the extension of the access time of the BL. Hence, ageing of the PG transistor is an important factor in SRAM write performance and affects the write performance negatively.

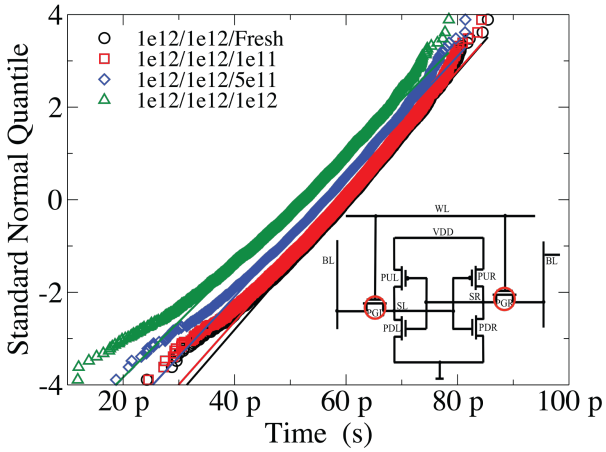


Fig.6.26 QQ plot of WM when PG transistors are degraded. The legend is the ageing levels of (PUR and PDL)/(PUL and PDR)/PG transistors [109].

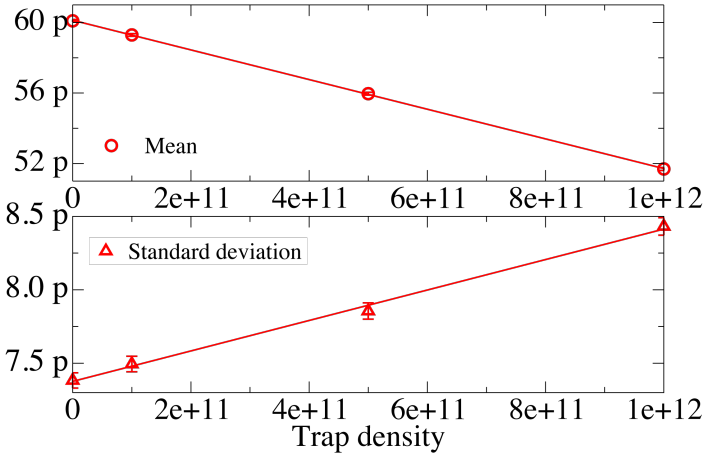


Fig.6.27 Evolution of the average WM and its standard deviation when PG transistors are degraded [109].

6.5 Response Surface

Previous simulation results and discussions show ageing effects of different transistors on SRAM stability and write performance. It illustrates that SNM and WM are both determined by the ageing of more than one transistor and that it is useful to analyse the response of the whole cell in order to understand how SNM and WM varies when different transistors in the cell suffer from different levels of degradation. Fig.6.28 shows a response surface of SNM and Fig.6.29 shows a response surface when a '1' is written to the SL side. In both figures, the X and Y axes represent the trap density of PDL & PUR and PUL & PDR transistors, respectively. PG transistors are kept fresh when PU and PD transistors' trap density is below $5x10^{11}cm^{-2}$. When PU and PD transistors' trap

density is the same as or above $5x10^{11}cm^{-2}$, PG transistors' trap density changes to $1x10^{11}cm^{-2}$. The response surface gives a direct view of the change of SNM and WM when two inverters age at different levels. The worst case for SNM shown in Fig.6.28 is when the two inverters have the maximum mismatch. The worst case for WM shown in Fig.6.29 is when the two inverters do not have any mismatch. Actually, this is the middle case because mismatch improves one side of the write performance and degrades the opposite side. Therefore, the worst case for WM is also when the two inverters have the maximum mismatch. This will guide the circuit designers to change the cell ratio in order to meet the design criteria.

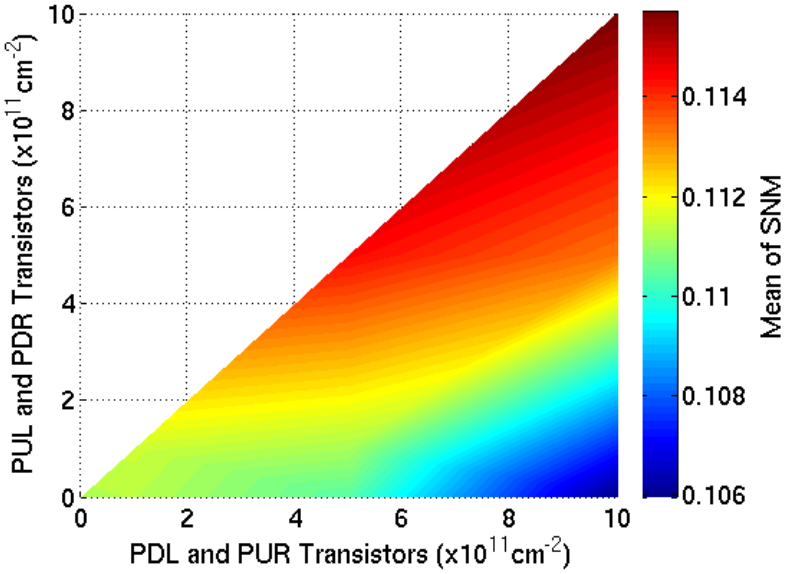


Fig.6.28 Response surface of SNM.

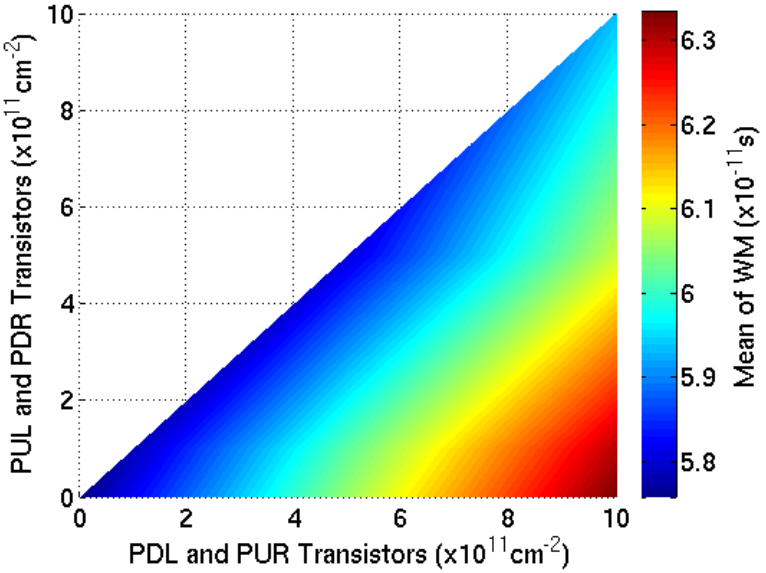


Fig.6.29 Response surface when '1' is written to the SL side. [109]

6.6 Summary

Using accurate reliability physical simulations, compact model extraction and compact model generation, we carried out the statistical circuit simulations of statistical variability and ageing effects, and thoroughly investigated the influence of BTI-induced transistor degradation on 20 nm bulk CMOS SRAM stability and dynamic write performance. For the first time, we have gained systematic understanding for the impact of ageing on the SRAM operation. To summarise, the transistor ageing increases the threshold voltage and weakens each transistor (the access, pull-up and pull-down abilities for PG, PU, PD transistors respectively). This can have both negative and positive impact on SRAM stability and the dynamic WM distribution depending on the different ageing scenarios for the different SRAM cell transistors. The different scenarios simulated in the work correspond to different SRAM operation conditions and read/write/hold patterns, affecting differently the cell's stability and write performance. In this study we found that the ageing of PUR (the PU transistor holding '1' before the write) and PG transistors have a great influence on WM. Ageing of PU transistors which hold '1' before the write makes the PG transistor easier to overcome that particular PU transistor for writing '1'. Ageing of PG transistors extend the access time of the BL and consequently reduce the write speed and increase the cell's stability. The mismatch between the two inverters results in the imbalance of the cell. It reduces the cell's stability and enlarges the

difference of writing time of ‘0’ and ‘1’, by improving one side of the write performance and degrading the opposite side. Response surface clearly shows the changes of SNM and WM in response to the change of different ageing levels between the two inverters.

Overall, the BTI based reliability analysis of the SRAM cell is workload and metric dependent. It is down to the designer to make the correct assumptions when analysing SRAM power, performance and yield as a function of degradation. Incorrect assumptions can lead to drastic overdesign on significant yield loss over time.

Chapter 7

Conclusions and future work

7.1 Summary and Conclusion

Very large-scale circuits, containing billions of transistors due to the rapid development of the IC industry, enable complex functions to be realized on a fingernail-sized chip. However, aggressive scaling makes it more and more difficult for every transistor to reach the specific performance target, resulting in instability of the circuit performance. As device dimensions scale to atomic level, statistical variability becomes one of the dominant factors accounting for this problem, making each transistor microscopically different. Additionally, BTI-induced traps do not only interact with statistical variability and give rise to the variations of device performance, but also shorten the device lifetime and thus the circuit lifetime.

Traditional modelling and characterization of a single device behavior and using it at circuit level is no longer a feasible option. Moreover, the customary circuit simulation method is no longer suitable to investigate the statistical circuit behavior due to the stochastic nature of statistical variability and BTI. For this reason, this research focuses on the statistical compact model extraction and generation methodologies, with which sufficiently large compact models can be generated, to accurately describe the statistical device behaviours. This research also focuses on the investigation of the statistical circuit behaviour under the influence of statistical variability and ageing, using SRAM as the simulation vehicle. The aim of this PhD is to integrate the impact of statistical variability and reliability into compact models and investigate its influence on SRAM performance

in advanced and emerging CMOS technology Based on the results and the data presented in the chapters above, the following conclusions can be drawn:

1. Device performance under statistical variability and BTI-induced ageing is obtained and analysed. Physical simulations have been performed with an ensemble of 1,000 devices at the ageing levels of 0 (containing statistical variability only), low, medium and high (containing statistical variability and at trap density of $1x10^{11}cm^{-2}$, $5x10^{11}cm^{-2}$, $1x10^{12}cm^{-2}$, respectively). Physical simulation results are used for the investigation of device performance, as well as the input data for the compact model extraction. Simulation results are analysed in 1) the statistical level, focusing on device performance distributions and the corresponding moments analysis within a large sample, 2) individual transistor level, focusing on the link between transistor performance shift and transistor's structure. For an individual device, trapping of a single discrete charge in the vicinity of a current "percolation path" can result in a significant change in transistor characteristics, including a large threshold voltage shift and on-current reduction. The maximum V_{TH} increase and on-current reduction are 145.3mV and 319μA/μm respectively. With ageing level increases, the mean value of V_{TH} increases (compared with fresh devices, the increase step is 3.08%, 15.48%, 31.22% at low, medium and high ageing levels respectively), accompanied with the decrease of I_{ON} (the decrease step is -0.69%, -3.91%, -7.67% compared with fresh devices) and I_{OFF} (the decrease step is -6.58%, -28.76%, -47.94%). The standard deviations of each figures of merit are increased as ageing level increases due to the larger performance variations ageing brings to the device.

2. Statistical compact models are successfully extracted against the physical simulation results with nearly all errors below 6%. Physical simulations give 1,000 devices' performance at each ageing level. Using the figures of merit based compact model parameter selection strategy and the two-stage compact model extraction method, the same number of compact models is extracted from physical simulations at each ageing level. In the second stage of the compact model extraction, seven key parameters are successfully selected to capture individual transistor performance under statistical variability and BTI-induced

ageing. They are VTH0, EAT0, VOFF, NFACTOR, UA, VSAT and CDSCD. The extraction results show that within the minimum choice of the model parameters, the extracted compact models are capable of modelling the behavior of variable devices at different ageing levels and capturing the correlations between device figures of merit. The first four moments of the selected parameters are approximately linearly dependent on trap density. This is then used with the combination of the GLD method to generate compact models at arbitrary trap density.

3. Compact models are successfully generated at arbitrary trap densities using the combination of the GLD and interpolation methods. The compact models generated using traditional Gaussian V_T method, are shown first and the inaccuracy stimulates the requirement of the accurate compact model generation approach. Using the result of the analysis in conclusion 2, that the first four moments of the re-extracted parameters are linearly dependent on trap density, each parameter's first four moments at any arbitrary trap density can be obtained through interpolation. Using the moment information as an input for the GLD method, compact models are successfully generated at an arbitrary trap density. An ageing model is also established translating between trap densities and ageing time. The compact model generation methodologies are successfully embedded into the SPICE simulator RandomSpice. The accuracy of the compact model generation approach and the ageing model is validated by comparing the generated compact models at 3 month and physical simulations at trap density of $7.5x10^{11}cm^{-2}$, which shows a high agreement. By generating 100,000 devices, it is verified the GLD method can follow the trend of the physical simulation and is more accurate than the Gaussian V_T method.

4. The influences of statistical variability and ageing on SRAM are evaluated by using the compact model generation approach. In the 6-T SRAM cell, the variation of each transistor's performance to the cell's stability and the speed of write performance are tested through the sensitivity test. In SNM test, it shows that the SRAM cell's stability is most sensitive to PD transistors, followed by the PG and PU transistors. In WM test, devices in the order from the highest to the

lowest sensitivity are PGR PUR, PGL, PDL, PUL, PDR (when writing '1' to the left side which holds '0' previously). Three scenarios are performed to fully investigate different ageing situations. Ageing induced mismatch in the two inverters makes the cell unstable, and enlarges the writing time difference between writing '0' and '1'. The ageing of PG transistors reduces the access time to the internal inverters. Thus the cell's stability is increased and the speed of writing is reduced. We have also found that in write performance, the PU transistor, which holds '1' before writing performance, plays the most important role between the two inverters. Ageing of this PU transistor enables a faster speed of the writing performance because it is easier for the PG transistor to overcome this PU transistor. The response surfaces for SNM and WM are shown as well, giving a very direct view of the change of SNM and WM as the two inverters are at different ageing levels.

7.2 Future Work

There are several possibilities to extend this work. However we suggest three main possible research areas below:

1) Very large simulations can be performed for the prediction of the yield. The yield can be predicted on large-scale SRAM circuits containing millions of transistors. With the compact model generation method used in this thesis, rare devices' influence can be explored through high sigma investigation. This will lead to the prediction of circuit life span.

2) The extraction and generation methodology can be used to other future CMOS technology and new structures. GLD method is based on the first four moments of the parameter values. If these values can be obtained accurately, GLD method is capable of generating compact models for other dimensional devices or new structure devices.

3) The investigation of BTI-induced ageing process in this work considers the static traps. The trapping and de-trapping procedures are not utilised in this work. Therefore, in

the future work, trapping and de-trapping methodology (dynamic trapping) can be incorporated and implemented into the circuit simulator.

Appendix A

Appendix A.1 V_{TH} at different trap densities for NMOS, V_{DS}=1V.

V_{TH} (V)						
Trap density (cm^{-2})	Mean	Standard deviation	Skewness	Kurtosis	Min	Max
Fresh	0.1137	0.0777	0.031	3.0	-0.1484	0.3569
1e11	0.1172	0.0782	0.057	3.0	-0.1273	0.3704
5e11	0.1313	0.0789	0.061	3.0	-0.1098	0.3888
1e12	0.1492	0.0805	0.045	3.0	-0.1071	0.3966

Appendix A.2 I_{ON} at different trap densities for NMOS, V_{DS}=1V.

I_{ON} (μA/μm)						
Trap density (cm^{-2})	Mean	Standard deviation	Skewness	Kurtosis	Min	Max
Fresh	1303	167	0.036	3.0	779	1826
1e11	1294	168	0.002	3.0	778	1797
5e11	1252	168	0.011	3.0	724	1765
1e12	1203	170	0.004	3.0	677	1730

Appendix A.3 I_{OFF} at different trap densities for NMOS, V_{DS}=1V.

$Log_{10}(I_{OFF})$						
Trap density (cm^{-2})	Mean	Standard deviation	Skewness	Kurtosis	Min	Max
Fresh	-6.593	0.830	-0.24	2.8	-9.2915	-4.4836
1e11	-6.630	0.840	-0.25	2.9	-9.6469	-4.5542
5e11	-6.778	0.854	-0.23	2.9	-9.8090	-4.6783
1e12	-6.968	0.880	-0.20	2.9	-9.8981	-4.6866

Appendix A.4 DIBL at different trap densities for NMOS, V_{DS}=1V.

DIBL(V/V)

Trap density (cm^{-2})	Mean	Standard deviation	Skewness	Kurtosis	Min	Max
Fresh	0.09401	0.0268	0.68	3.5	0.03448	0.20247
1e11	0.09423	0.0268	0.65	3.4	0.03356	0.20247
5e11	0.09461	0.0271	0.61	3.4	0.03260	0.19829
1e12	0.09496	0.0277	0.64	3.6	0.03500	0.20717

Appendix A.5 V_{TH} at different trap densities for NMOS, V_{DS}=0.05V.

V_{TH} (V)						
Trap density (cm^{-2})	Mean	Standard deviation	Skewness	Kurtosis	Min	Max
Fresh	0.2077	0.0708	0.071	2.99	-0.0336	0.4323
1e11	0.2115	0.0716	0.107	2.96	0.0042	0.4408
5e11	0.2259	0.0725	0.092	2.96	0.0085	0.4498
1e12	0.2441	0.0742	0.073	2.93	0.0282	0.4731

Appendix A.6 I_{ON} at different trap densities for NMOS, V_{DS}=0.05V.

I_{ON} (μA/μm)						
Trap density (cm^{-2})	Mean	Standard deviation	Skewness	Kurtosis	Min	Max
Fresh	218	16.4	0.030	3.03	167	274
1e11	217	16.5	-0.001	3.04	163	273
5e11	214	16.8	-0.028	3.05	159	270
1e12	211	17.2	-0.087	3.11	152	270

Appendix A. 7 $Log_{10}(I_{OFF})$ at different trap densities for NMOS, V_{DS}=0.05V.

$Log_{10}(I_{OFF})$						
Trap density (cm^{-2})	Mean	Standard deviation	Skewness	Kurtosis	Min	Max
Fresh	-7.522	0.770	-0.123	2.89	-9.907	-5.135
1e11	-7.561	0.779	-0.150	2.90	-10.250	-5.434
5e11	-7.712	0.791	-0.127	2.89	-10.342	-5.474
1e12	-7.903	0.812	-0.102	2.88	-10.431	-5.633

Appendix A.8 V_{TH} at different trap densities for PMOS, V_{DS}=1V.

V_{TH} (V)						
Trap density (cm^{-2})	Mean	Standard deviation	Skewness	Kurtosis	Min	Max
Fresh	-0.1123	0.0778	-0.011	3.02	0.1009	-0.3696
1e11	-0.1166	0.0782	-0.020	2.99	0.0957	-0.3736
5e11	-0.1311	0.0795	-0.017	3.03	0.0888	-0.3958
1e12	-0.1479	0.0810	-0.016	3.07	0.0875	-0.4340

Appendix A.9 I_{ON} at different trap densities for PMOS, V_{DS}=1V.

I_{ON} (μA/μm)						
Trap density (cm^{-2})	Mean	Standard deviation	Skewness	Kurtosis	Min	Max
Fresh	979	125	-0.004	2.91	632	1345
1e11	971	125	-0.010	2.90	599	1345
5e11	941	127	-0.024	2.93	553	1339
1e12	905	127	-0.005	2.89	473	1299

Appendix A.10 $Log_{10}(I_{OFF})$ at different trap densities for PMOS, V_{DS}=1V.

$Log_{10}(I_{OFF})$						
Trap density (cm^{-2})	Mean	Standard deviation	Skewness	Kurtosis	Min	Max
Fresh	-6.552	0.818	-0.25	2.89	-9.295	-4.709
1e11	-6.597	0.826	-0.24	2.85	-9.338	-4.723
5e11	-6.745	0.848	-0.22	2.86	-9.531	-4.729
1e12	-6.921	0.874	-0.19	2.85	-9.967	-4.794

Appendix A.11 DIBL at different trap densities for PMOS.

DIBL (V/V)						
Trap density (cm^{-2})	Mean	Standard deviation	Skewness	Kurtosis	Min	Max
Fresh	-0.1111	0.0265	-0.52	3.31	-0.0450	-0.2208

1e11	-0.1113	0.0268	-0.53	3.31	-0.0443	-0.2207
5e11	-0.1115	0.0274	-0.53	3.31	-0.0414	-0.2283
1e12	-0.1120	0.0281	-0.52	3.26	-0.0414	-0.2286

Appendix A.12 V_{TH} at different trap densities for PMOS, V_{DS}=0.05V.

V_{TH} (V)						
Trap density (cm^{-2})	Mean	Standard deviation	Skewness	Kurtosis	Min	Max
Fresh	-0.2235	0.0684	-0.15	3.03	-0.0359	-0.4544
1e11	-0.2280	0.0689	-0.15	3.00	-0.0359	-0.4631
5e11	-0.2425	0.0702	-0.14	3.00	-0.0552	-0.4746
1e12	-0.2599	0.0721	-0.14	3.02	-0.0552	-0.5144

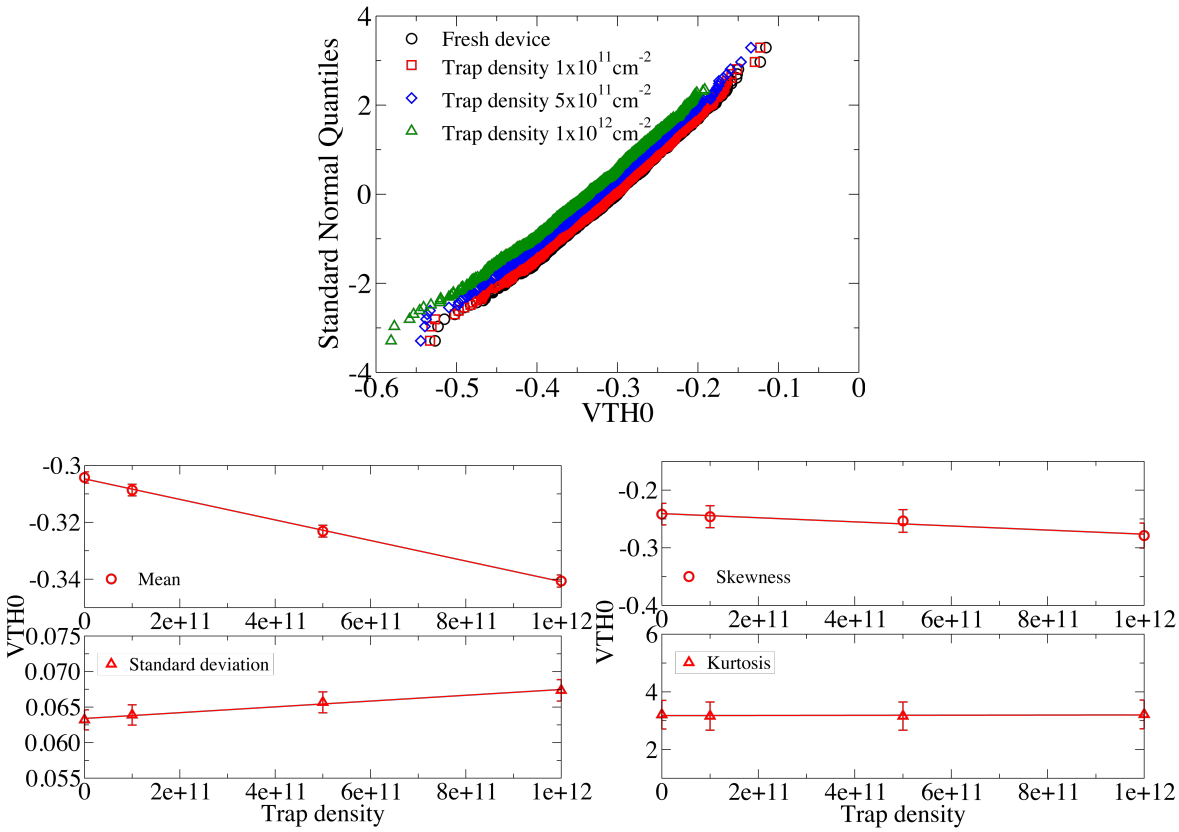
Appendix A.13 I_{ON} at different trap densities for PMOS, V_{DS}=0.05V.

I_{ON} (μA/μm)						
Trap density (cm^{-2})	Mean	Standard deviation	Skewness	Kurtosis	Min	Max
Fresh	122.7	8.49	-0.12	2.84	96.9	145.6
1e11	122.3	8.52	-0.11	2.84	96.9	145.4
5e11	120.7	8.66	-0.12	2.81	94.4	144.2
1e12	118.8	8.80	-0.10	2.85	87.0	143.4

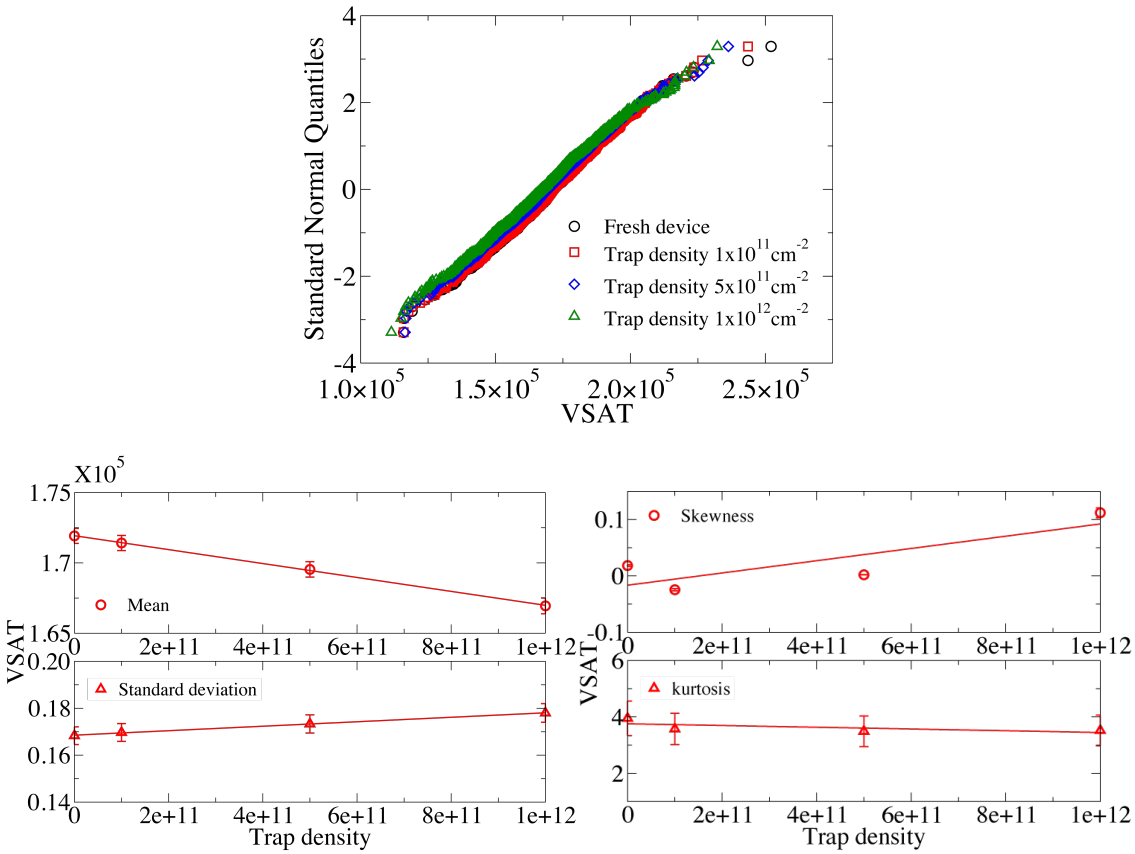
Appendix A.14 $Log_{10}(I_{OFF})$ at different trap densities for PMOS, V_{DS}=0.05V.

$Log_{10}(I_{OFF})$						
Trap density (cm^{-2})	Mean	Standard deviation	Skewness	Kurtosis	Min	Max
Fresh	-7.062	0.748	-0.18	2.92	-10.016	-5.652
1e11	-7.662	0.753	-0.17	2.89	-10.050	-5.652
5e11	-7.813	0.769	-0.15	2.91	-10.365	-5.860
1e12	-7.995	0.793	-0.14	2.91	-10.800	-5.860

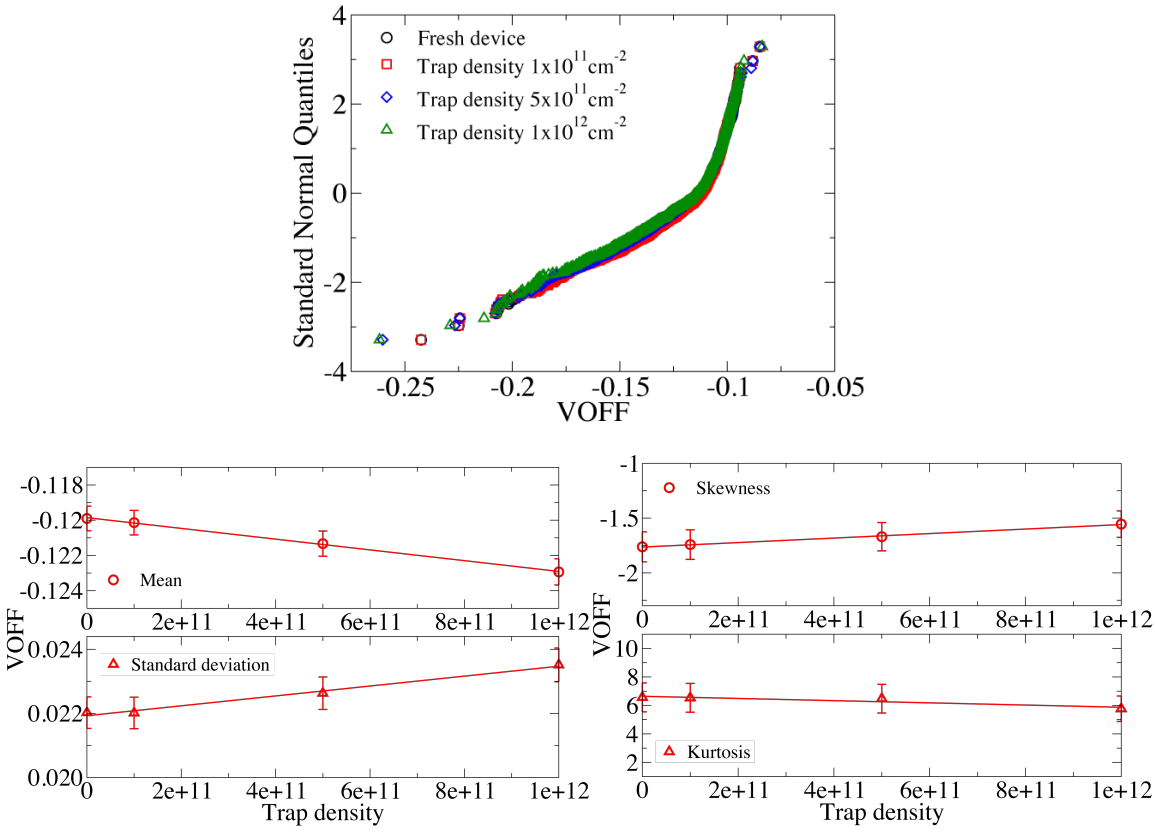
Appendix B



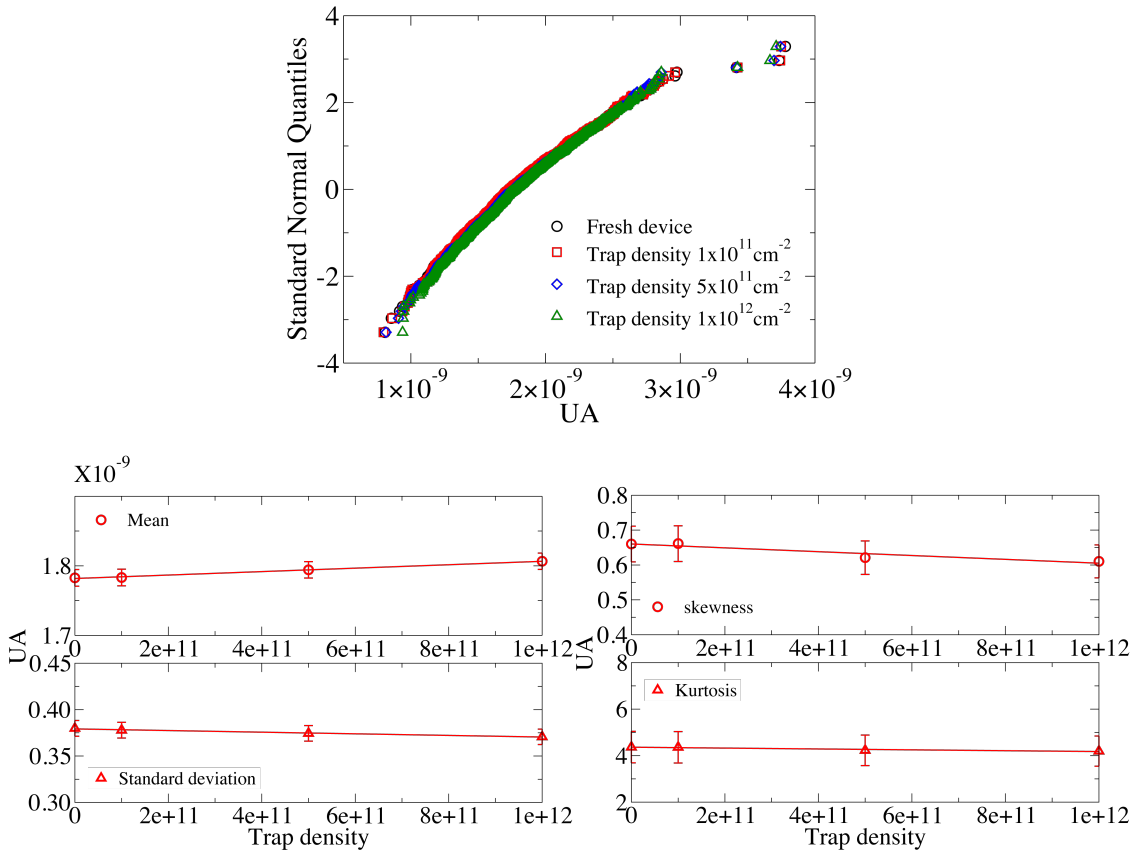
Appendix B.1 The distribution of VTH0 for PMOS.



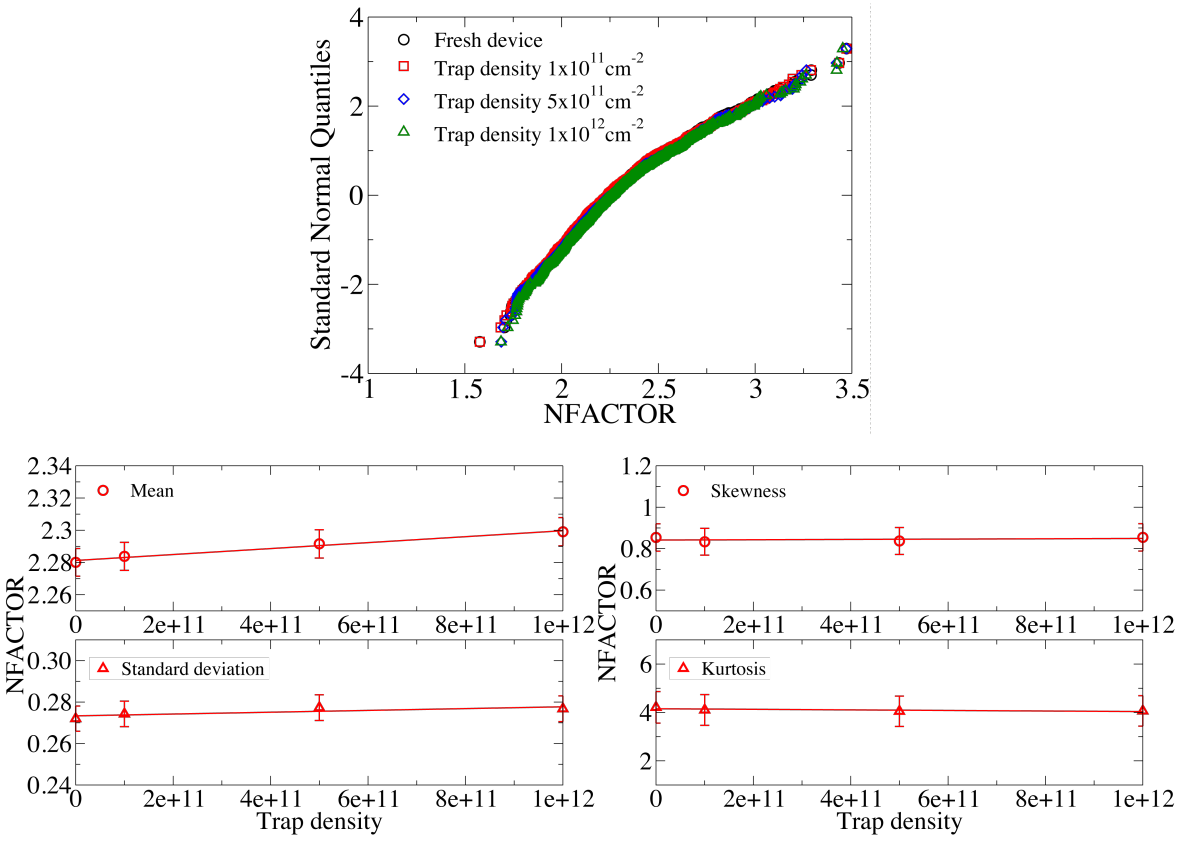
Appendix B.2 The distribution of VSAT for PMOS.



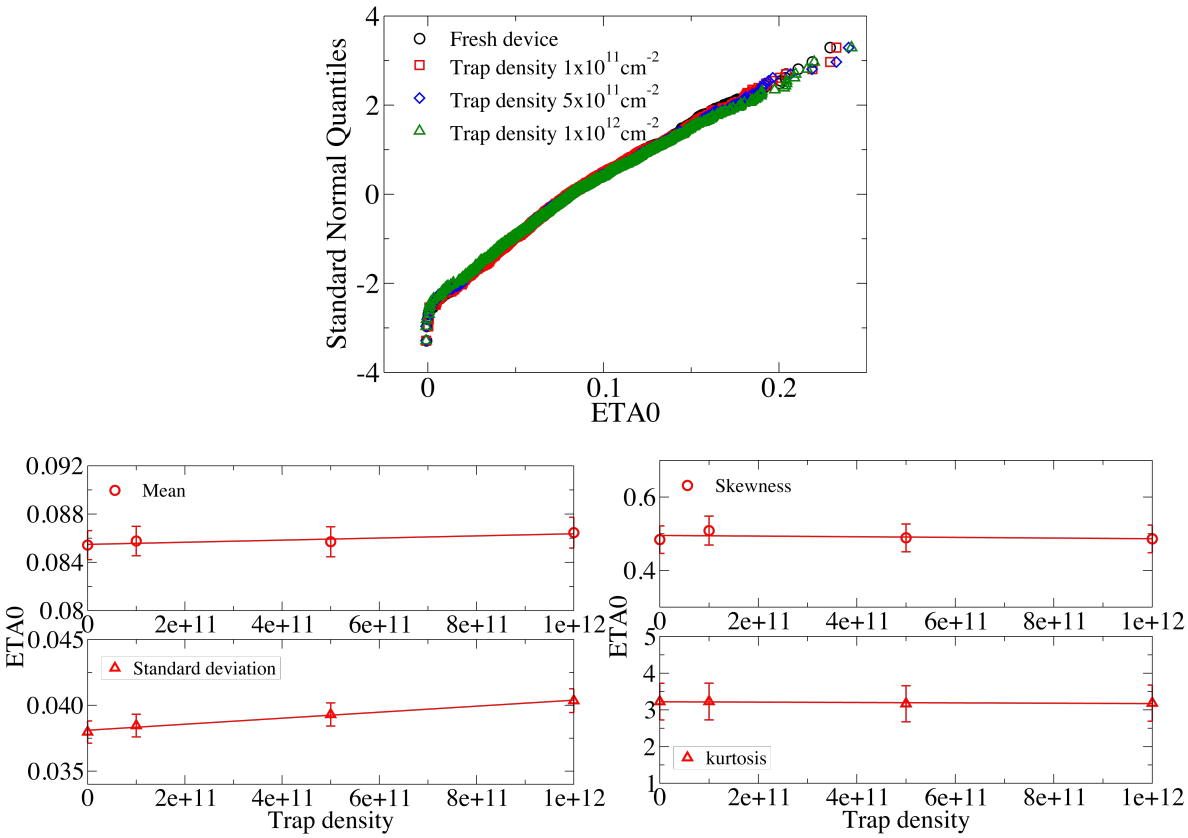
Appendix B.3 The distribution of VOFF for PMOS.



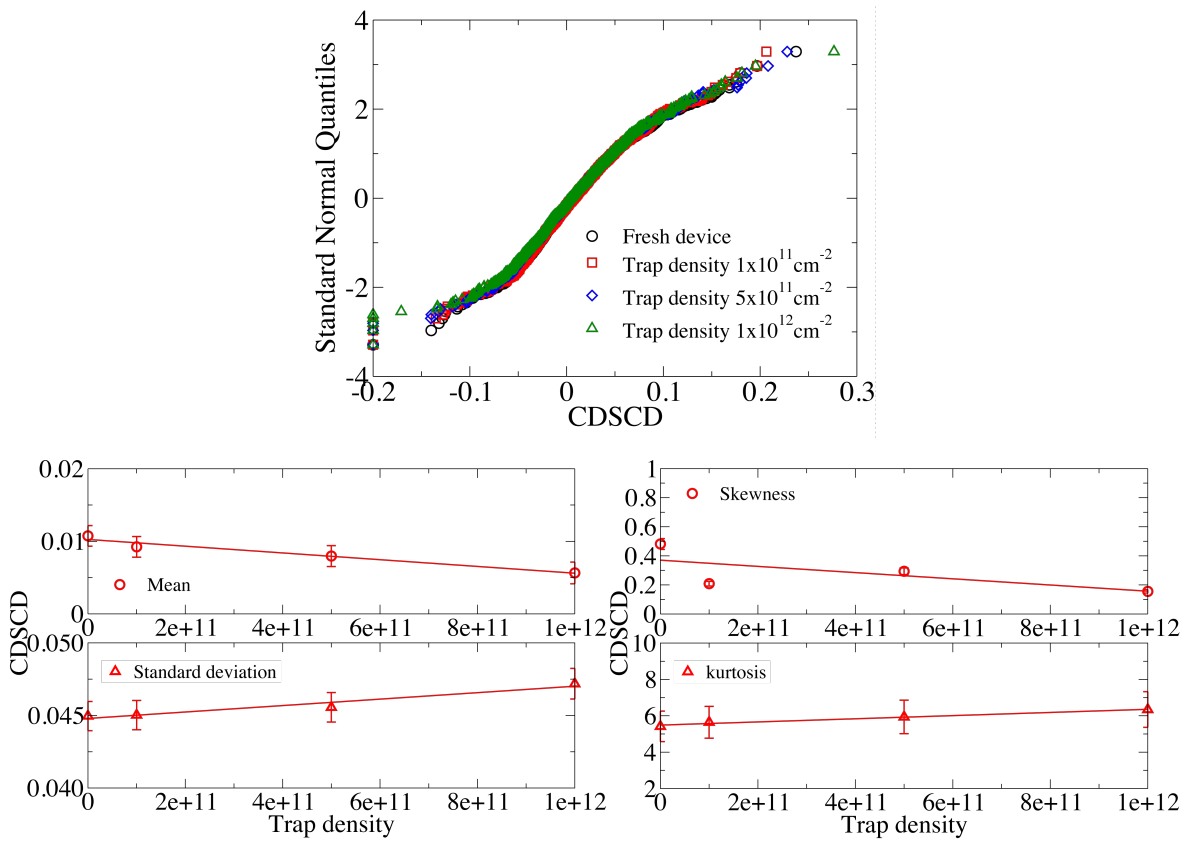
Appendix B.4 The distribution of UA for PMOS.



Appendix B.5 The distribution of NFACTOR for PMOS.

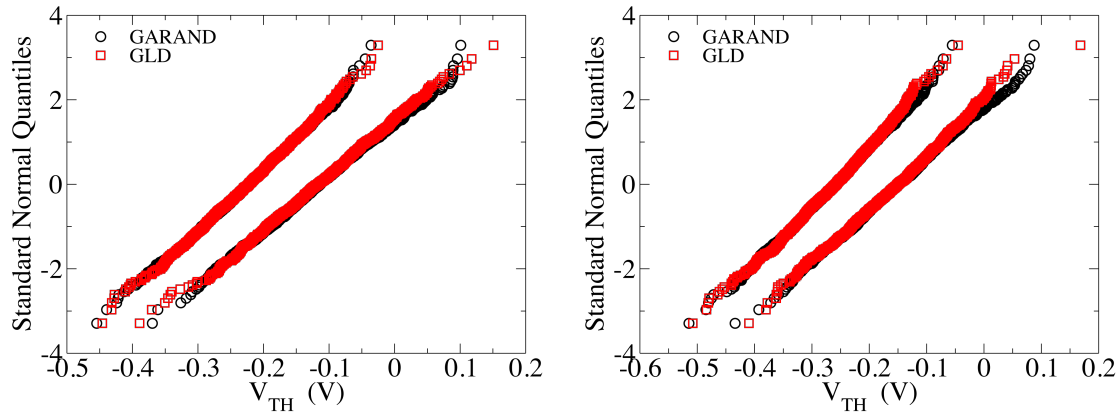


Appendix B.6 The distribution of ETA0 for PMOS.

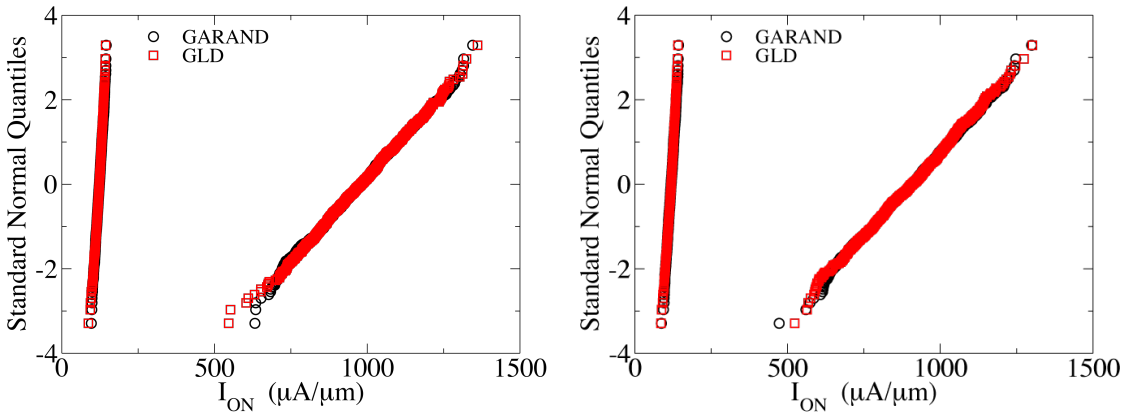


Appendix B.7 The distribution of CDSCD for PMOS.

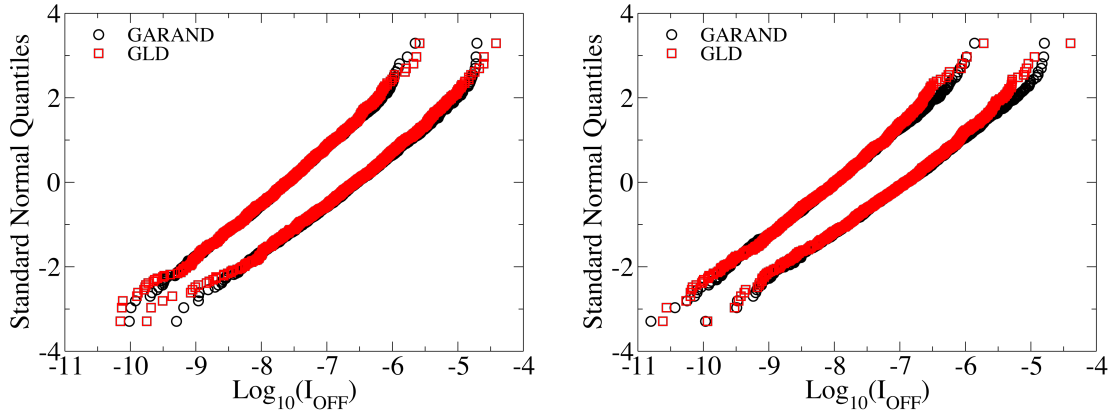
Appendix C



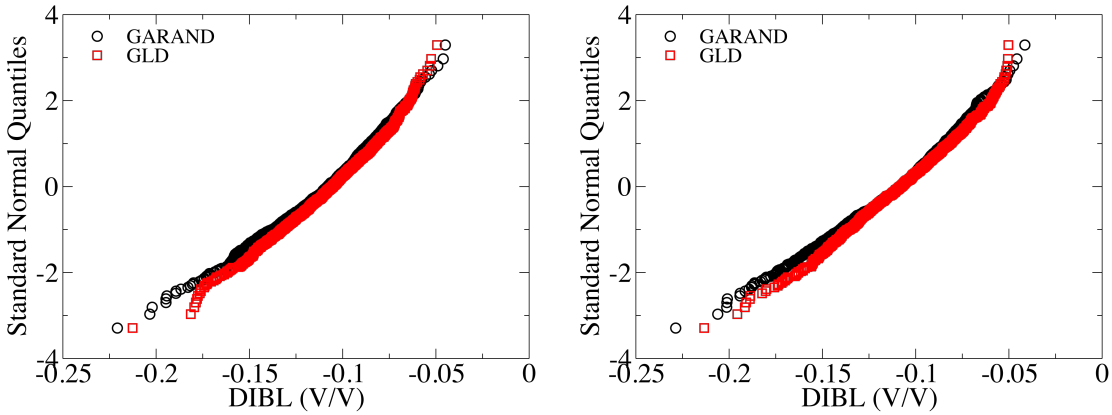
Appendix C.1 Comparisons of V_{TH} between physical simulation and GLD generated compact models for PMOS devices. The left and right figures are at trap density of 0 and $1x10^{12}cm^{-2}$ respectively.



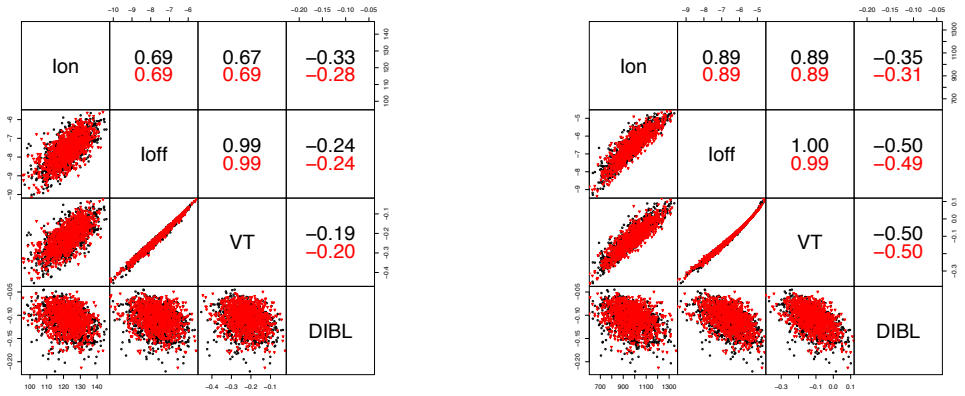
Appendix C.2 Comparisons of I_{ON} between physical simulation and GLD generated compact models for PMOS devices. The left and right figures are at trap density of 0 and $1x10^{12}cm^{-2}$ respectively.



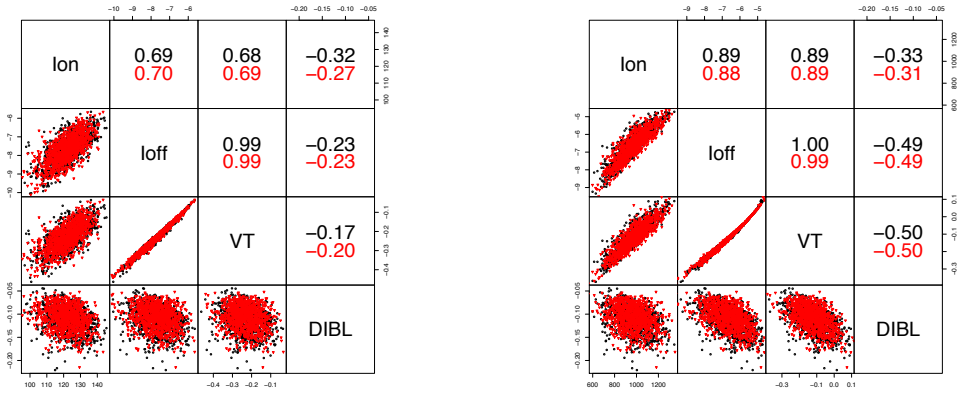
Appendix C.3 Comparisons of I_{OFF} between physical simulation and GLD generated compact models for PMOS devices. The left and right figures are at trap density of 0 and $1x10^{12}cm^{-2}$ respectively.



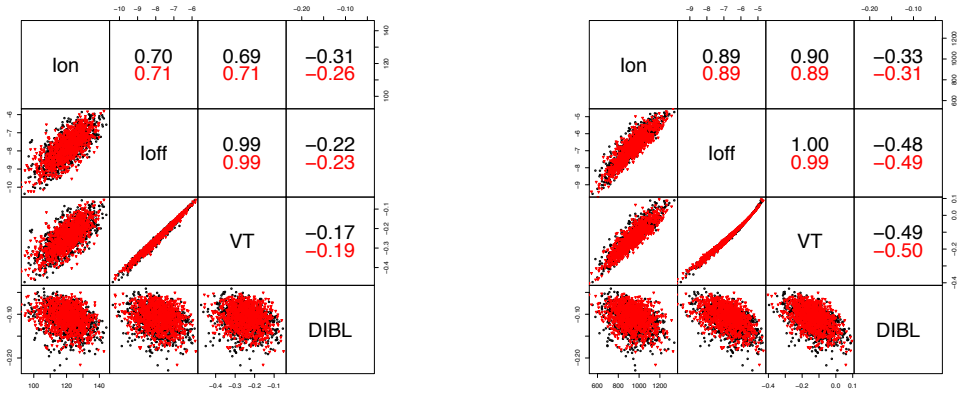
Appendix C.4 Comparisons of DIBL between physical simulation and GLD generated compact models for PMOS devices. The left and right figures are at trap density of 0 and $1x10^{12}cm^{-2}$ respectively.



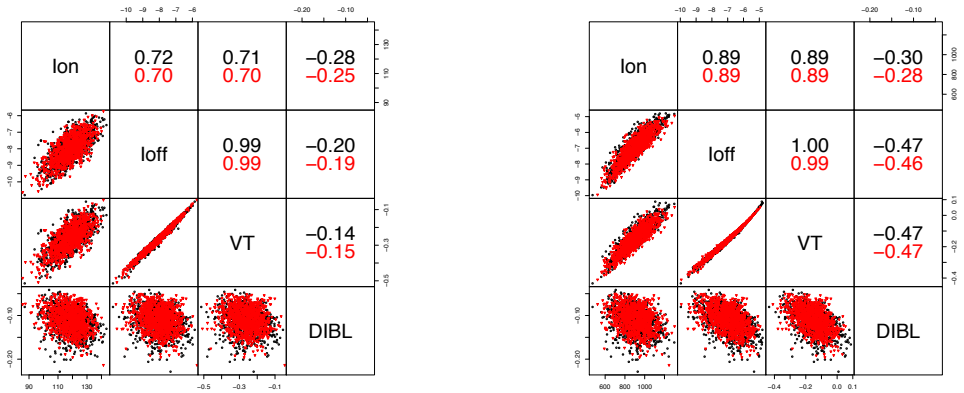
Appendix C.5 Correlations of figures of merit of PMOS between physical simulation and GLD generated compact models at trap density of 0. The left figure is when V_{DS}=0.05V, while the right figure is when V_{DS}=1V.



Appendix C.6 Correlations of figures of merit of PMOS between physical simulation and GLD generated compact models at trap density of $1x10^{11}cm^{-2}$. The left figure is when V_{DS}=0.05V, while the right figure is when V_{DS}=1V.



Appendix C.7 Correlations of figures of merit of PMOS between physical simulation and GLD generated compact models at trap density of $5x10^{11}cm^{-2}$. The left figure is when V_{DS}=0.05V, while the right figure is when V_{DS}=1V.



Appendix C.8 Correlations of figures of merit of PMOS between physical simulation and GLD generated compact models at trap density of $1x10^{12}cm^{-2}$. The left figure is when V_{DS}=0.05V, while the right figure is when V_{DS}=1V.

Bibliography

[1] Y. Chauhan, S. Venugopalan, M. A. Karim, S. Khandelwal, N. Paydavosi, P. Thakur*, et al.*, "BSIM - Industry standard compact MOSFET models," in *Solid-State Device Research Conference (ESSDERC), 2012 Proceedings of the European*, 2012, pp. 46-49.

[2] . *http://www-device.eecs.berkeley.edu/bsim/?page=BSIM4*.

[3] A. Cathignol, B. Cheng, D. Chanemougame, A. R. Brown, K. Rochereau, G. Ghibaudo*, et al.*, "Quantitative evaluation of statistical variability sources in a 45-nm-technological node LP N-MOSFET," *Ieee Electron Device Letters,* vol. 29, pp. 609-611, Jun 2008.

[4] H. P. Tuinhout, A. H. Montree, J. Schmitz, and P. A. Stolk, "Effects of gate depletion and boron penetration on matching of deep submicron CMOS transistors," *International Electron Devices Meeting - 1997, Technical Digest,* pp. 631-634, 1997.

[5] C. Auth, A. Cappellani, J. S. Chun, A. Dalis, A. Davis, T. Ghani*, et al.*, "45nm High-k + metal gate strain-enhanced transistors," in *VLSI Technology, 2008 Symposium on*, 2008, pp. 128-129.

[6] S. Pae, M. Agostinelli, M. Brazie, R. Chau, G. Dewey, T. Ghani*, et al.*, "BTI reliability of 45 nm high-k plus metal-gate process technology," *2008 Ieee International Reliability Physics Symposium Proceedings - 46th Annual,* pp. 352-357, 2008.

[7] J. S. Kilby, "Invention of the integrated circuit," *Electron Devices, IEEE Transactions on,* vol. 23, pp. 648-654, 1976.

[8] H. Iwai, "CMOS scaling for sub-90 nm to sub-10 nm," in *VLSI Design, 2004. Proceedings. 17th International Conference on*, 2004, pp. 30-35.

[9] X. Sun, "Nanoscale Bulk MOSFET Design and Process Technology for Reduced Variability," Doctor of Philosophy, Electrical Engineering and Computer Sciences, University of California, Berkeley, 2010.

[10] G. E. Moore, "cramming more components onto integrated circuits," *Electronics,* vol. 38, April 19, 1965 1965.

[11] S. Borkar, "Design challenges of technology scaling," *Ieee Micro,* vol. 19, pp. 23-29, Jul-Aug 1999.

[12] P. P. Gelsinger, "Microprocessors for the new millennium: Challenges, opportunities, and new frontiers," in *Solid-State Circuits Conference, 2001. Digest of Technical Papers. ISSCC. 2001 IEEE International*, 2001, pp. 22-25.

[13] M. Bohr, "A 30 Year Retrospective on Dennard's MOSFET Scaling Paper," *Solid-State Circuits Society Newsletter, IEEE,* vol. 12, pp. 11-13, 2007.

[14] B. Hoeneisen and C. A. Mead, "Fundamental limitations in microelectronics—I. MOS technology," *Solid-State Electronics,* vol. 15, pp. 819-829, 7// 1972.

[15] G. Declerck, "A look into the future of nanoelectronics," *2005 Symposium on VLSI Technology, Digest of Technical Papers,* pp. 6-10, 2005.

[16] C. M. Mezzomo, A. Bajolet, A. Cathignol, E. Josse, and G. Ghibaudo, "Modeling local electrical fluctuations in 45 nm heavily pocket-implanted bulk MOSFET," *Solid-State Electronics,* vol. 54, pp. 1359-1366, Nov 2010.

[17] N. Z. Haron and S. Hamdioui, "Why is CMOS scaling coming to an END?," in *Design and Test Workshop, 2008. IDT 2008. 3rd International*, 2008, pp. 98-103.

[18] M. Bohr, "The new era of scaling in an SoC world," in *Solid-State Circuits Conference - Digest of Technical Papers, 2009. ISSCC 2009. IEEE International*, 2009, pp. 23-28.

[19] . *http://www.intel.com/pressroom/kits/advancedtech/doodle/ref_HiK-MG/high-k.htm*.

[20] A. Asenov, S. Roy, R. A. Brown, G. Roy, C. Alexander, C. Riddet*, et al.*, "Advanced simulation of statistical variability and reliability in nano CMOS transistors," in *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, 2008, pp. 1-1.

[21] R. Keyes, "The effect of randomness in the distribution of impurity atoms on FET thresholds," *Applied physics,* vol. 8, pp. 251-259, 1975/11/01 1975.

[22] G. Roy, A. R. Brown, F. Adamu-Lema, S. Roy, and A. Asenov, "Simulation study of individual and combined sources of intrinsic parameter fluctuations in conventional nano-MOSFETs," *Ieee Transactions on Electron Devices,* vol. 53, pp. 3063-3070, Dec 2006.

[23] K. Takeuchi, A. Nishida, and T. Hiramoto, "Random Fluctuations in Scaled MOS Devices," *International Conference on Simulation of Semiconductor Processes and Devices,* pp. 79-85, 2009.

[24] A. S. M. Zain, S. Markov, B. J. Cheng, and A. Asenov, "Comprehensive study of the statistical variability in a 22 nm fully depleted ultra-thin-body SOI MOSFET," *Solid-State Electronics,* vol. 90, pp. 51-55, Dec 2013.

[25] H.-S. P. Wong, Y. Taur, and D. J. Frank, "Discrete random dopant distribution e ffects in nanometer-scale MOSFETs," *Microelectronics Reliability,* vol. 38, 1998.

[26] K. Nagase, S. I. Ohkawa, M. Aoki, and H. Masuda, "Variation status in 100nm CMOS process and below," *Icmts 2004: Proceedings of the 2004 International Conference on Microelectronic Test Structures,* pp. 257-261, 2004.

[27] N. Seoane, G. Indalecio, E. Comesana, M. Aldegunde, A. J. Garcia-Loureiro, and K. Kalna, "Random Dopant, Line-Edge Roughness, and Gate Workfunction Variability in a Nano InGaAs FinFET," *Electron Devices, IEEE Transactions on,* vol. 61, pp. 466-472, 2014.

[28] W. Liping, A. R. Brown, C. Millar, A. Burenkov, W. Xingsheng, A. Asenov*, et al.*, "Simulation for statistical variability in realistic 20nm MOSFET," in *Ultimate Integration on Silicon (ULIS), 2014 15th International Conference on*, 2014, pp. 5-8.

[29] A. Asenov, F. Adamu-Lema, X. S. Wang, and S. M. Amoroso, "Problems With the Continuous Doping TCAD Simulations of Decananometer CMOS Transistors," *Ieee Transactions on Electron Devices,* vol. 61, pp. 2745-2751, Aug 2014.

[30] W. Xingsheng, D. Reid, W. Liping, A. Burenkov, C. Millar, B. Cheng*, et al.*, "Variability-aware compact model strategy for 20-nm bulk MOSFETs," in *Simulation of Semiconductor Processes and Devices (SISPAD), 2014 International Conference on*, 2014, pp. 293-296.

[31] T. R. David, "Large-Scale Simulations of Intrinsic Parameter Fluctuations in Nano-Scale MOSFETs," PhD, Department of Electronics and Electrical Engineering, University of Glasgow, University of Glasgow, 2010.

[32] D. Burnett, K. Erington, C. Subramanian, and K. Baker, "Implications of Fundamental Threshold Voltage Variations for High-Density Sram and Logic Circuits," *1994 Symposium on Vlsi Technology,* pp. 15-16, 1994.

[33] T. Mizuno, J. Okumtura, and A. Toriumi, "Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant number in MOSFET's," *Electron Devices, IEEE Transactions on,* vol. 41, pp. 2216-2221, 1994.

[34] K. R. Lakshmikumar, R. A. Hadaway, and M. A. Copeland, "Characterisation and modeling of mismatch in MOS transistors for precision analog design," *Solid-State Circuits, IEEE Journal of,* vol. 21, pp. 1057-1066, 1986.

[35] J. T. Horstmann, U. Hilleringmann, and K. F. Goser, "Matching analysis of deposition defined 50-nm MOSFET's," *Electron Devices, IEEE Transactions on,* vol. 45, pp. 299-306, 1998.

[36] T. Hagiwara, K. Yamaguchi, and S. Asai, "Threshold Voltage Deviation in Very Small MOS Transistors Due to Local Impurity Fluctuations," in *VLSI Technology, 1982. Digest of Technical Papers. Symposium on*, 1982, pp. 46-47.

[37] M. Steyaert, J. Bastos, R. Roovers, P. Kinget, W. Sansen, B. Graindourze*, et al.*, "Threshold Voltage Mismatch in Short-Channel Mos-Transistors," *Electronics Letters,* vol. 30, pp. 1546-1548, Sep 1 1994.

[38] H. S. P. Wong, Y. Taur, and D. J. Frank, "Discrete random dopant distribution effects in nanometer-scale MOSFETs," *Microelectronics and Reliability,* vol. 38, pp. 1447-1456, Sep 1998.

[39] A. Asenov, "Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 μm MOSFET's: A 3-D “atomistic” simulation study," *Electron Devices, IEEE Transactions on,* vol. 45, pp. 2505-2513, 1998.

[40] A. R. Brown and A. Asenov, "Capacitance fluctuations in bulk MOSFETs due to random discrete dopants," *Journal of Computational Electronics,* vol. 7, pp. 115-118, Sep 2008.

[41] K. Nishinohara, N. Shigyo, and T. Wada, "Effects of Microscopic Fluctuations in Dopant Distributions on Mosfet Threshold Voltage," *Ieee Transactions on Electron Devices,* vol. 39, pp. 634-639, Mar 1992.

[42] P. A. Stolk and D. B. M. Klaassen, "The effect of statistical dopant fluctuations on MOS device performance," *Iedm - International Electron Devices Meeting, Technical Digest 1996,* pp. 627-630, 1996.

[43] K. J. Kuhn, "Reducing variation in advanced logic technologies: Approaches to process and design for manufacturability of nanoscale CMOS," *2007 Ieee International Electron Devices Meeting, Vols 1 and 2,* pp. 471-474, 2007.

[44] T. Linton, M. Chandhok, B. J. Rice, and G. Schrom, "Determination of the line edge roughness specification for 34 nm devices," *International Electron Devices 2002 Meeting, Technical Digest,* pp. 303-306, 2002.

[45] Y. Ban, S. Sundareswaran, R. Panda, and D. Z. Pan, "Electrical Impact of Line-Edge Roughness on Sub-45nm Node Standard Cell," *Design for Manufacturability through Design-Process Integration Iii,* vol. 7275, 2009.

[46] T. Linton, M. Giles, and P. Packan, "The impact of Line Edge Roughness on 100nm Device Performance," presented at the Silicon Nanoelectronics Workshop, Kyoto, Japan, 1999.

[47] T. D. Linton, S. F. Yu, and R. Shaheed, "3D modelling of fluctuation effects in highly scaled VLSI devices," *Vlsi Design,* vol. 13, pp. 103-109, 2001.

[48] P. Oldiges, Q. Lin, K. Petrillo, M. Sanchez, I. Meikei, and M. Hargrove, "Modeling line edge roughness effects in sub 100 nanometer gate length devices," in *Simulation of Semiconductor Processes and Devices, 2000. SISPAD 2000. 2000 International Conference on*, 2000, pp. 131-134.

[49] J. Wu, J. H. Chen, and K. P. Liu, "Transistor width dependence of LER degradation to CMOS device characteristics," *Sispad 2002: International Conference on Simulation of Semiconductor Processes and Devices,* pp. 95-98, 2002.

[50] A. Asenov, S. Kaya, and A. R. Brown, "Intrinsic parameter fluctuations in decananometer MOSFETs introduced by gate line edge roughness," *Ieee Transactions on Electron Devices,* vol. 50, pp. 1254-1260, May 2003.

[51] A. R. Brown, N. M. Idris, J. R. Watling, and A. Asenov, "Impact of Metal Gate Granularity on Threshold Voltage Variability: A Full-Scale Three-Dimensional Statistical Simulation Study," *Ieee Electron Device Letters,* vol. 31, pp. 1199-1201, Nov 2010.

[52] H. Dadgour, K. Endo, V. De, and K. Banerjee, "Modeling and Analysis of Grain-Orientation Effects in Emerging Metal-Gate Devices and Implications for SRAM Reliability," *Ieee International Electron Devices Meeting 2008, Technical Digest,* pp. 705-708, 2008.

[53] C. H. Hwang, T. Y. Li, M. H. Han, K. F. Lee, H. W. Cheng, and Y. M. Li, "Statistical Analysis of Metal Gate Workfunction Variability, Process Variation, and Random Dopant Fluctuation in Nano-CMOS Circuits," *2009 International Conference on Simulation of Semiconductor Processes and Devices,* pp. 99-102, 2009.

[54] X. A. Zhang, J. Li, M. Grubbs, M. Deal, B. Magyari-Kope, B. M. Clemens*, et al.*, "Physical Model of the Impact of Metal Grain Work Function Variability on Emerging Dual Metal Gate MOSFETs and its Implication for SRAM Reliability," *2009 Ieee International Electron Devices Meeting,* pp. 51-54, 2009.

[55] G. Indalecio, N. Seoane, M. Aldegunde, K. Kalna, and A. J. Garcia-Loureiro, "Scaling of Metal Gate Workfunction Variability in nanometer SOI-FinFETs," *2014 15th International Conference on Ultimate Integration on Silicon (Ulis),* pp. 105-108, 2014.

[56] M. J. Cho, J. D. Lee, M. Aoulaiche, B. Kaczer, P. Roussel, T. Kauerauf*, et al.*, "Insight Into N/PBTI Mechanisms in Sub-1-nm-EOT Devices," *Ieee Transactions on Electron Devices,* vol. 59, pp. 2042-2048, Aug 2012.

[57] H. Amrouch, J. Martin-Martinez, V. M. van Santen, M. Moras, R. Rodriguez, M. Nafria*, et al.*, "Connecting the physical and application level towards grasping aging effects," in *Reliability Physics Symposium (IRPS), 2015 IEEE International*, 2015, pp. 3D.1.1-3D.1.8.

[58] B. Kaczer, J. Franco, M. Toledano-Luque, P. J. Roussel, M. F. Bukhori, A. Asenov*, et al.*, "The Relevance of Deeply-Scaled FET Threshold Voltage Shifts for Operation Lifetimes," *2012 Ieee International Reliability Physics Symposium (Irps),* 2012.

[59] M. F. Bukhori, S. Roy, and A. Asenov, "Statistical aspects of reliability in bulk MOSFETs with multiple defect states and random discrete dopants," *Microelectronics Reliability,* vol. 48, pp. 1549-1552, Aug-Sep 2008.

[60] A. R. Brown, V. Huard, and A. Asenov, "Statistical Simulation of Progressive NBTI Degradation in a 45-nm Technology pMOSFET," *Ieee Transactions on Electron Devices,* vol. 57, pp. 2320-2323, Sep 2010.

[61] J. X. Fang and S. S. Sapatnekar, "Understanding the Impact of Transistor-Level BTI Variability," *2012 Ieee International Reliability Physics Symposium (Irps),* 2012.

[62] C. C. Chen, S. Cha, T. Z. Liu, and L. Milor, "System-Level Modeling of Microprocessor Reliability Degradation Due to BTI and HCI," *2014 Ieee International Reliability Physics Symposium,* 2014.

[63] D. K. Schroder and J. A. Babcock, "Negative bias temperature instability: Road to cross in deep submicron silicon semiconductor manufacturing," *Journal of Applied Physics,* vol. 94, pp. 1-18, Jul 1 2003.

[64] N. Goel, P. Dubey, J. Kawa, and S. Mahapatra, "Impact of time-zero and NBTI variability on sub-20nm FinFET based SRAM at low voltages," in *Reliability Physics Symposium (IRPS), 2015 IEEE International*, 2015, pp. CA.5.1-CA.5.7.

[65] . *https://www.si2.org/cmc_index.php*.

[66] B. J. Sheu, D. L. Scharfetter, P. K. Ko, and M. C. Jeng, "Bsim - Berkeley Short-Channel Igfet Model for Mos-Transistors," *Ieee Journal of Solid-State Circuits,* vol. 22, pp. 558-566, Aug 1987.

[67] http://www-device.eecs.berkeley.edu/bsim/?page=BSIM4.

[68] I. Agbo, M. Taouil, S. Hamdioui, H. Kukner, P. Weckx, P. Raghavan*, et al.*, "Integral impact of BTI and voltage temperature variation on SRAM sense amplifier," in *VLSI Test Symposium (VTS), 2015 IEEE 33rd*, 2015, pp. 1-6.

[69] S. Drapatz, "Parametric Reliability of 6T-SRAM Core Cell Arrays," PhD, Electronics, Technical University of Munich, Munich, 2011.

[70] S. Drapatz, "Parametric Reliability of 6T-SRAM Core Cell Arrays," PhD dissertation, Department of Electronic Engineering, Technical University of Munich, Munich, 2011.

[71] S. Khan, I. Agbo, S. Hamdioui, H. Kukner, B. Kaczer, P. Raghavan*, et al.*, "Bias Temperature Instability analysis of FinFET based SRAM cells," *2014 Design, Automation and Test in Europe Conference and Exhibition (Date),* 2014.

[72] A. Asenov, S. Roy, R. A. Brown, G. Roy, C. Alexander, C. Riddet*, et al.*, "Advanced simulation of statistical variability and reliability in nano CMOS

transistors," *Ieee International Electron Devices Meeting 2008, Technical Digest,* pp. 421-421, 2008.

[73] P. Weckx, B. Kaczer, P. J. Roussel, F. Catthoor, and G. Groeseneken, "Impact of time-dependent variability on the yield and performance of 6T SRAM cells in an advanced HK/MG technology," in *IC Design & Technology (ICICDT), 2015 International Conference on*, 2015, pp. 1-4.

[74] (03/01/2015). *http://www.goldstandardsimulations.com/products/*.

[75] S. Selberherr, "Mos Device Modeling at 77-K," *Ieee Transactions on Electron Devices,* vol. 36, pp. 1464-1474, Aug 1989.

[76] T. Ando, A. B. Fowler, and F. Stern, "Electronic-Properties of Two-Dimensional Systems," *Reviews of Modern Physics,* vol. 54, pp. 437-672, 1982.

[77] U. Ravaioli, "Hierarchy of simulation approaches for hot carrier transport in deep submicron devices," *Semiconductor Science and Technology,* vol. 13, pp. 1-10, Jan 1998.

[78] Z. Ren, R. Venugopal, S. Datta, and M. Lundstrom, "Examination of design and manufacturing issues in a 10 nm double gate MOSFET using nonequilibrium Green's function simulation," in *Electron Devices Meeting, 2001. IEDM '01. Technical Digest. International*, 2001, pp. 5.4.1-5.4.4.

[79] M. G. Ancona and G. J. Iafrate, "Quantum Correction to the Equation of State of an Electron-Gas in a Semiconductor," *Physical Review B,* vol. 39, pp. 9536-9540, May 1 1989.

[80] A. R. Brown, A. Asenov, and J. R. Watling, "Intrinsic fluctuations in sub 10-nm double-gate MOSFETs introduced by discreteness of charge and matter," *Ieee Transactions on Nanotechnology,* vol. 1, pp. 195-200, Dec 2002.

[81] . *http://www.goldstandardsimulations.com/products/garand/*.

[82] W. F. J. Frank and B. S. Berry, "Lattice Location and Atomic Mobility of Implanted Boron in Silicon," *Radiation Effects and Defects in Solids,* vol. 21, pp. 105-111, 1974.

[83] C. K. Birdsall and D. Fuss, "Clouds-in-clouds, clouds-in-cells physics for many-body plasma simulation (Reprinted from the Journal of Computational Physics, vol 3, pg 494-511, 1969)," *Journal of Computational Physics,* vol. 135, pp. 141-148, Aug 1997.

[84] R. W. Hockney and J. W. Eastwood, *Computer Simulations Using Particles*.

[85] T. Masaharu, O. Tohru, and O. Naofumi, "A New Algorithm for Threw-Dimensional Voronoi Tessellation " *Journal of Computational Physics,* vol. 51, pp. 191-207, 1983.

[86] . *http://www.goldstandardsimulations.com/products/mystic/*.

[87] C. Binjie, D. Dideban, N. Moezi, C. Millar, G. Roy, W. Xingsheng*, et al.*, "Statistical-Variability Compact-Modeling Strategies for BSIM4 and PSP," *Design & Test of Computers, IEEE,* vol. 27, pp. 26-35, 2010.

[88] "BSIM4 MOSFET Model User's Manual."

[89] P. Asenov, "Accurate Statistical Circuit Simulation in the Presence of Statistical Variability," PhD dissertation, School of Engineering, University of Glasgow, Glasgow, 2013.

[90] N. Moezi, "Statistical compact model strategies for nano CMOS transistors subject of atomic scale variability," dissertation, Dept. Electron. Eng, University of Glasgow, Glasgow, 2012.

[91] N. Moezi, D. Dideban, B. J. Cheng, S. Roy, and A. Asenov, "Impact of statistical parameter set selection on the statistical compact model accuracy: BSIM4 and PSP case study," *Microelectronics Journal,* vol. 44, pp. 7-14, Jan 2013.

[92] J. X. Fang and S. S. Sapatnekar, "The Impact of BTI Variations on Timing in Digital Logic Circuits," *Ieee Transactions on Device and Materials Reliability,* vol. 13, pp. 277-286, Mar 2013.

[93] S. Pae, J. Maiz, C. Prasad, and B. Woolery, "Effect of BTI Degradation on Transistor Variability in Advanced Semiconductor Technologies," *Ieee Transactions on Device and Materials Reliability,* vol. 8, pp. 519-525, Sep 2008.

[94] R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte Carlo Method*: WILEY, 2011.

[95] B. G. Tabashnick and L. S. Fidell, *Using Multivariate Statistics*, 2006.

[96] P. Chaparala, J. Shibley, and P. Lim, "Threshold voltage drift in PMOSFETS due to NBTI and HCI," *2000 Ieee International Integrated Reliability Workshop Final Report,* pp. 95-97, 2000.

[97] B. Cheng, N. Moezi, D. Dideban, G. Roy, S. Roy, and A. Asenov, "Benchmarking the Accuracy of PCA Generated Statistical Compact Model Parameters Against Physical Device Simulation and Directly Extracted Statistical

Parameters," *2009 International Conference on Simulation of Semiconductor Processes and Devices,* pp. 143-146, 2009.

[98] C. Hastings, F. Mosteller, J. W. Tukey, and C. P. Winsor, "Low Moments for Small Samples - a Comparative Study of Order Statistics," *Annals of Mathematical Statistics,* vol. 18, pp. 413-426, 1947.

[99] S. R. John and B. W. Schmeiser, "An approximate method for generating asymmetric random variables," *Communications of the ACM,* vol. 17, pp. 78-82, 1974.

[100] F. Marshall, K. Georgia, S. M. Govind, and L. Thomas, "a study of the generalized tukey lambda family," *Communications in Statistics - Theory and Methods,* vol. 17, pp. 3547-3567, 1988.

[101] S. Steve, "A Discretized Approach to Flexibly Fit Generalized Lambda Distributions to Data," *Journal of Modern Applied Statistical Methods,* vol. 4, pp. 408-424, 2005.

[102] J. Ding, D. Reid, M. Campbell, and A. Asenov, "An accurate compact modelling approach for statistical ageing and reliability," in *Simulation of Semiconductor Processes and Devices (SISPAD), 2013 International Conference on*, 2013, pp. 57-60.

[103] M. Toledano-Luque, B. Kaczer, J. Franco, P. J. Roussel, T. Grasser, T. Y. Hoffmann*, et al.*, "From mean values to distributions of BTI lifetime of deeply scaled FETs through atomistic understanding of the degradation," in *VLSI Technology (VLSIT), 2011 Symposium on*, 2011, pp. 152-153.

[104] M. Toledano-Luque, B. Kaczer, J. Franco, P. J. Roussel, T. Grasser, and G. Groeseneken, "Defect-centric perspective of time-dependent BTI variability," *Microelectronics Reliability,* vol. 52, pp. 1883-1890, Sep-Oct 2012.

[105] A. Pushkarna, S. Raghavan, and H. Mahmoodi, "Comparison of performance parameters of SRAM designs in 16nm CMOS and CNTFET technologies," in *SOC Conference (SOCC), 2010 IEEE International*, 2010, pp. 339-342.

[106] P. Bai, C. Auth, S. Balakrishnan, M. Bost, R. Brain, V. Chikarmane*, et al.*, "A 65nm logic technology featuring 35nm gate lengths, enhanced channel strain, 8 Cu interconnect layers, low-k ILD and 0.57 μm² SRAM cell," in *Electron Devices Meeting, 2004. IEDM Technical Digest. IEEE International*, 2004, pp. 657-660.

[107] R. Rogenmoser and L. T. Clark, "Reducing Transistor Variability for Higher-Performance, Lower-Power Chips," *Ieee Micro,* vol. 33, pp. 18-26, Mar-Apr 2013.

[108] E. Seevinck, F. J. List, and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," *Solid-State Circuits, IEEE Journal of,* vol. 22, pp. 748-754, 1987.

[109] J. Ding, D. Reid, P. Asenov, C. Millar, and A. Asenov, "Influence of Transistors With BTI-Induced Aging on SRAM Write Performance," *Ieee Transactions on Electron Devices,* vol. 62, pp. 3133-3138, Oct 2015.

[110] T. Y. Chan, J. Chen, P. K. Ko, and C. Hu, "The impact of gate-induced drain leakage current on MOSFET scaling," in *Electron Devices Meeting, 1987 International*, 1987, pp. 718-721.